Text Classification and Naïve Bayes

The Task of Text Classification
Is this spam?

Subject: Important notice!
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients;;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.


Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

© Stanford University. All Rights Reserved.
Who wrote which Federalist papers?

• 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
• Authorship of 12 of the letters in dispute
• 1963: solved by Mosteller and Wallace using Bayesian methods
Male or female author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam...(Male)

2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...(Female)

Positive or negative movie review?

- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.
What is the subject of this article?

MEDLINE Article

MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

Dan Jurafsky
Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- ...

Dan Jurafsky
Text Classification: definition

- **Input:**
  - a document $d$
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

- **Output:** a predicted class $c \in C$
Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive
Classification Methods: Supervised Machine Learning

• **Input:**
  - a document \( d \)
  - a fixed set of classes \( C = \{c_1, c_2, \ldots, c_J\} \)
  - A training set of \( m \) hand-labeled documents \( (d_1, c_1), \ldots, (d_m, c_m) \)

• **Output:**
  - a learned classifier \( \gamma : d \rightarrow c \)
Classification Methods: Supervised Machine Learning

• Any kind of classifier
  • Naïve Bayes
  • Logistic regression
  • Support-vector machines
  • k-Nearest Neighbors

• ... of course, neural net/deep learning!!
Text Classification and Naïve Bayes

Naïve Bayes
Naïve Bayes Intuition

• Simple ("naïve") classification method based on Bayes rule
• Relies on very simple representation of document
  • Bag of words
The bag of words representation

<table>
<thead>
<tr>
<th></th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>great</td>
<td>2</td>
</tr>
<tr>
<td>love</td>
<td>2</td>
</tr>
<tr>
<td>recommend</td>
<td>1</td>
</tr>
<tr>
<td>laugh</td>
<td>1</td>
</tr>
<tr>
<td>happy</td>
<td>1</td>
</tr>
</tbody>
</table>

\( \gamma ( ) ) = c \)
Bag of words for document classification

Test document

parser
language
label
translation

Machine Learning

parser
training
algorithm
shrinkage
network...

NLP

garbage
collection
memory
optimization
region...

Planning

planning
temporal
reasoning
plan
language...

GUI

...
Naïve Bayes Classifier (I)

$$c_{MAP} = \arg\max_{c \in C} P(c \mid d)$$

$$= \arg\max_{c \in C} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \arg\max_{c \in C} P(d \mid c)P(c)$$

Dropping the denominator

MAP is “maximum a posteriori” = most likely class
Naïve Bayes Classifier (II)

\[ c_{MAP} = \arg \max_{c \in C} P(d \mid c)P(c) \]

\[ = \arg \max_{c \in C} P(x_1, x_2, \ldots, x_n \mid c)P(c) \]

Document d represented as features x1..xn
Naïve Bayes Classifier (IV)

\[ c_{MAP} = \arg \max_{c \in C} P(x_1, x_2, \ldots, x_n | c)P(c) \]

- \( O(|X|^n \cdot |C|) \) parameters
- Could only be estimated if a very, very large number of training examples was available.
- How often does this class occur?
- We can just count the relative frequencies in a corpus.
Multinomial Naïve Bayes Independence Assumptions

\[ P(x_1, x_2, \ldots, x_n \mid c) \]

- **Bag of Words assumption**: Assume position doesn’t matter
- **Conditional Independence**: Assume the feature probabilities \( P(x_i \mid c_j) \) are independent given the class \( c \).

\[ P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \cdot P(x_2 \mid c) \cdot P(x_3 \mid c) \cdot \ldots \cdot P(x_n \mid c) \]
Multinomial Naïve Bayes Classifier

\[ c_{MAP} = \arg \max_{c \in C} P(x_1, x_2, \ldots, x_n \mid c)P(c) \]

\[ c_{NB} = \arg \max_{c \in C} P(c_j) \prod_{x \in X} P(x \mid c) \]
Applying Multinomial Naive Bayes Classifiers to Text Classification

positions ← all word positions in test document

\[ c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{i \in \text{positions}} P(x_i | c_j) \]
Learning the Multinomial Naïve Bayes Model

- First attempt: maximum likelihood estimates
  - simply use the frequencies in the data

\[
\hat{P}(c_j) = \frac{\text{doccount}(C = c_j)}{N_{\text{doc}}}
\]

\[
\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}
\]
Parameter estimation

\[ \hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)} \]

fraction of times word \( w_i \) appears among all words in documents of topic \( c_j \)

- Create mega-document for topic \( j \) by concatenating all docs in this topic
- Use frequency of \( w \) in mega-document
Problem with Maximum Likelihood

- What if we have seen no training documents with the word *fantastic* and classified in the topic *positive (thumbs-up)*?

\[
\hat{P}(\text{"fantastic"} | \text{positive}) = \frac{\text{count("fantastic", positive)}}{\sum_{w \in V} \text{count}(w, \text{positive})} = 0
\]

- Zero probabilities cannot be conditioned away, no matter the other evidence!

\[
c_{\text{MAP}} = \arg\max_c \hat{P}(c) \prod_i \hat{P}(x_i | c)
\]
Generative Model for Multinomial Naïve Bayes

c=China

X₁=Shanghai  X₂=and  X₃=Shenzhen  X₄=issue  X₅=bonds
Naïve Bayes and Language Modeling

- Naïve bayes classifiers can use any sort of feature
  - URL, email address, dictionaries, network features
- But if, as in the previous slides
  - We use only word features
  - we use all of the words in the text (not a subset)
- Then
  - Naïve bayes has an important similarity to language modeling.
Each class = a unigram language model

- Assigning each word: $P(\text{word} \mid c)$
- Assigning each sentence: $P(s \mid c) = \prod P(\text{word} \mid c)$

Class $pos$

<table>
<thead>
<tr>
<th>pos</th>
<th>l</th>
<th>love</th>
<th>this</th>
<th>fun</th>
<th>film</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>love</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>this</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>fun</td>
<td>0.1</td>
<td>0.1</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>0.1</td>
<td>film</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$P(s \mid pos) = 0.0000005$
Naïve Bayes as a Language Model

- Which class assigns the higher probability to $s$?

<table>
<thead>
<tr>
<th></th>
<th>Model pos</th>
<th></th>
<th>Model neg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.2</td>
<td>__________</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>I</td>
<td>love</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.001</td>
<td>love</td>
</tr>
<tr>
<td></td>
<td>love</td>
<td>love</td>
<td>this</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.01</td>
<td>this</td>
</tr>
<tr>
<td></td>
<td>this</td>
<td>this</td>
<td>Fun</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.005</td>
<td>fun</td>
</tr>
<tr>
<td></td>
<td>fun</td>
<td>fun</td>
<td>film</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>film</td>
</tr>
<tr>
<td>P(s</td>
<td>pos)</td>
<td>&gt;</td>
<td>P(s</td>
</tr>
</tbody>
</table>

$P(s|\text{pos}) > P(s|\text{neg})$
Choosing a class:

\[ P(c \mid d5) \propto \frac{3}{4} \times \left(\frac{3}{7}\right)^3 \times \frac{1}{14} \times \frac{1}{14} \approx 0.0003 \]

\[ P(j \mid d5) \propto \frac{1}{4} \times (\frac{2}{9})^3 \times \frac{2}{9} \times \frac{2}{9} \approx 0.0001 \]
Naïve Bayes in Spam Filtering

- SpamAssassin Features:
  - Mentions Generic Viagra
  - Online Pharmacy
  - Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
  - Phrase: impress ... girl
  - From: starts with many numbers
  - Subject is all capitals
  - HTML has a low ratio of text to image area
  - One hundred percent guaranteed
  - Claims you can be removed from the list
  - 'Prestigious Non-Accredited Universities'
  - [http://spamassassin.apache.org/tests_3_3_x.html](http://spamassassin.apache.org/tests_3_3_x.html)
Summary: Naive Bayes is Not So Naive

• Very Fast, low storage requirements

• Robust to Irrelevant Features
  Irrelevant Features cancel each other without affecting results

• Very good in domains with many equally important features
  Decision Trees suffer from *fragmentation* in such cases – especially if little data

• Optimal if the independence assumptions hold: If assumed independence is correct, then it is the Bayes Optimal Classifier for problem

• A good dependable baseline for text classification
  • But we will see other classifiers that give better accuracy
Text Classification and Naïve Bayes

Precision, Recall, and the F measure
Precision and recall

- **Precision**: % of selected items that are correct
- **Recall**: % of correct items that are selected

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>
A combined measure: $F$

- A combined measure that assesses the P/R tradeoff is $F$ measure (weighted harmonic mean):

\[
F = \frac{1}{\frac{1}{P} + \frac{1}{R} \alpha \left(1 - \alpha\right)} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}
\]

- The harmonic mean is a very conservative average; see IIR § 8.3

- People usually use balanced F1 measure
  - i.e., with $\beta = 1$ (that is, $\alpha = \frac{1}{2}$):

\[
F = \frac{2PR}{P+R}
\]
More Than Two Classes: Sets of binary classifiers

- Dealing with *any-of* or *multivalue* classification
  - A document can belong to 0, 1, or >1 classes.

- For each class $c \in C$
  - Build a classifier $\gamma_c$ to distinguish $c$ from all other classes $c' \in C$

- Given test doc $d$,
  - Evaluate it for membership in each class using each $\gamma_c$
  - $d$ belongs to *any* class for which $\gamma_c$ returns true
More Than Two Classes: Sets of binary classifiers

- **One-of or multinomial classification**
  - Classes are mutually exclusive: each document in exactly one class

- For each class \( c \in C \)
  - Build a classifier \( \gamma_c \) to distinguish \( c \) from all other classes \( c' \in C \)

- Given test doc \( d \),
  - Evaluate it for membership in each class using each \( \gamma_c \)
  - \( d \) belongs to the one class with maximum score
Evaluation: Classic Reuters-21578 Data Set

- Most (over)used data set, 21,578 docs (each 90 types, 200 tokens)
- 9603 training, 3299 test articles (ModApte/Lewis split)
- 118 categories
  - An article can be in more than one category
  - Learn 118 binary category distinctions
- Average document (with at least one category) has 1.24 classes
- Only about 10 out of 118 categories are large

Common categories (#train, #test)

- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)
- Trade (369, 119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)
The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nation's pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter
**Confusion matrix c**

For each pair of classes \(c_1, c_2\) how many documents from \(c_1\) were incorrectly assigned to \(c_2\)?

- \(c_{3,2}\): 90 wheat documents incorrectly assigned to poultry

<table>
<thead>
<tr>
<th>Docs in test set</th>
<th>Assigned UK</th>
<th>Assigned poultry</th>
<th>Assigned wheat</th>
<th>Assigned coffee</th>
<th>Assigned interest</th>
<th>Assigned trade</th>
</tr>
</thead>
<tbody>
<tr>
<td>True UK</td>
<td>95</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>True poultry</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True wheat</td>
<td>10</td>
<td>90</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True coffee</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>34</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>True interest</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>True trade</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>14</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>
Per class evaluation measures

**Recall**: 
Fraction of docs in class $i$ classified correctly:

$$\frac{c_{ii}}{\sum_{j} c_{ij}}$$

**Precision**: 
Fraction of docs assigned class $i$ that are actually about class $i$:

$$\frac{c_{ii}}{\sum_{j} c_{ji}}$$

**Accuracy**: $(1 - \text{error rate})$
Fraction of docs classified correctly:

$$\frac{\sum_{i} c_{ii}}{\sum_{j} \sum_{i} c_{ij}}$$
Micro- vs. Macro-Averaging

• If we have more than one class, how do we combine multiple performance measures into one quantity?

• **Macroaveraging**: Compute performance for each class, then average.

• **Microaveraging**: Collect decisions for all classes, compute contingency table, evaluate.
**Micro- vs. Macro-Averaging: Example**

<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
<th>Micro Ave. Table</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Truth: yes</td>
<td>Truth: no</td>
</tr>
<tr>
<td>Classifier: yes</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Classifier: no</td>
<td>10</td>
<td>970</td>
</tr>
<tr>
<td>Classifier: yes</td>
<td>10</td>
<td>890</td>
</tr>
</tbody>
</table>

- Macroaveraged precision: \((0.5 + 0.9)/2 = 0.7\)
- Microaveraged precision: \(100/120 = .83\)
- Microaveraged score is dominated by score on common classes

Sec 15.2.4
Development Test Sets and Cross-validation

- Metric: P/R/F1 or Accuracy
- Unseen test set
  - avoid overfitting (‘tuning to the test set’)
  - more conservative estimate of performance
- Cross-validation over multiple splits
  - Handle sampling errors from different datasets
  - Pool results over each split
  - Compute pooled dev set performance
Accuracy as a function of data size

- With enough data
  - Classifier may not matter

Brill and Banko on spelling correction
How to tweak performance

• Domain-specific features and weights: very important in real performance

• Sometimes need to collapse terms:
  • Part numbers, chemical formulas, ...
  • But stemming generally doesn’t help

• Upweighting: Counting a word as if it occurred twice:
  • title words (Cohen & Singer 1996)
  • first sentence of each paragraph (Murata, 1999)
  • In sentences that contain title words (Ko et al, 2002)
Word Meaning and Similarity

Word Senses and Word Relations
Reminder: lemma and wordform

- A **lemma** or citation form
  - Same stem, part of speech, rough semantics
- A **wordform**
  - The “inflected” word as it appears in text

<table>
<thead>
<tr>
<th>Wordform</th>
<th>Lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>banks</td>
<td>bank</td>
</tr>
<tr>
<td>sung</td>
<td>sing</td>
</tr>
<tr>
<td>duermes</td>
<td>dormir</td>
</tr>
</tbody>
</table>
Lemmas have senses

• One lemma “bank” can have many meanings:
  • ...a **bank** can hold the investments in a custodial account...
  • “…as agriculture burgeons on the east **bank** the river will shrink even more”

**Sense 1:**

• Sense (or word sense)
  • A discrete representation of an aspect of a word’s meaning.

• The lemma **bank** here has two senses
Homonymy

**Homonyms**: words that share a form but have unrelated, distinct meanings:

- \( \text{bank}_1 \): financial institution, \( \text{bank}_2 \): sloping land
- \( \text{bat}_1 \): club for hitting a ball, \( \text{bat}_2 \): nocturnal flying mammal

1. Homographs (bank/bank, bat/bat)
2. Homophones:
   1. *Write* and *right*
   2. *Piece* and *peace*
Homonymy causes problems for NLP applications

- Information retrieval
  - “bat care”
- Machine Translation
  - bat: murciélago (animal) or bate (for baseball)
- Text-to-Speech
  - bass (stringed instrument) vs. bass (fish)
Polysemy

• 1. The **bank** was constructed in 1875 out of local red brick.
• 2. I withdrew the money from the **bank**
• Are those the same sense?
  • Sense 2: “A financial institution”
  • Sense 1: “The building belonging to a financial institution”
• A **polysemous** word has related meanings
  • Most non-rare words have multiple meanings
Metonymy or Systematic Polysemy: A systematic relationship between senses

- Lots of types of polysemy are systematic
  - School, university, hospital
  - All can mean the institution or the building.
- A systematic relationship:
  - Building ↔ Organization
- Other such kinds of systematic polysemy:
  - Author (Jane Austen wrote Emma)
  - Works of Author (I love Jane Austen)
  - Tree (Plums have beautiful blossoms)
  - Fruit (I ate a preserved plum)
How do we know when a word has more than one sense?

- The “zeugma” test: Two senses of serve?
  - Which flights serve breakfast?
  - Does Lufthansa serve Philadelphia?
  - Does Lufthansa serve breakfast and San Jose?

- Since this conjunction sounds weird,
  - we say that these are two different senses of “serve”
Synonyms

• Word that have the same meaning in some or all contexts.
  • filbert / hazelnut
  • couch / sofa
  • big / large
  • automobile / car
  • vomit / throw up
  • Water / H₂O

• Two lexemes are synonyms
  • if they can be substituted for each other in all situations
  • If so they have the same propositional meaning
Synonyms

• But there are few (or no) examples of perfect synonymy.
  • Even if many aspects of meaning are identical
  • Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.

• Example:
  • Water/H₂O
  • Big/large
  • Brave/courageous
Synonymy is a relation between senses rather than words

- Consider the words *big* and *large*
- Are they synonyms?
  - How *big* is that plane?
  - Would I be flying on a *large* or small plane?
- How about here:
  - Miss Nelson became a kind of *big* sister to Benjamin.
  - Miss Nelson became a kind of *large* sister to Benjamin.
- Why?
  - *big* has a sense that means being older, or grown up
  - *large* lacks this sense
Antonyms

- Senses that are opposites with respect to one feature of meaning
- Otherwise, they are very similar!
  - dark/light  short/long  fast/slow  rise/fall
  - hot/cold    up/down    in/out

- More formally: antonyms can
  - define a binary opposition
    - or be at opposite ends of a scale
      - long/short, fast/slow

- Be reversives:
  - rise/fall, up/down
Hyponymy and Hypernymy

• One sense is a **hyponym** of another if the first sense is more specific, denoting a subclass of the other
  • *car* is a hyponym of *vehicle*
  • *mango* is a hyponym of *fruit*

• Conversely **hypernym/superordinate** (“hyper is super”)
  • *vehicle* is a hypernym of *car*
  • *fruit* is a hypernym of *mango*

<table>
<thead>
<tr>
<th>Superordinate/ hyper</th>
<th>vehicle</th>
<th>fruit</th>
<th>furniture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subordinate/ hyponym</td>
<td>car</td>
<td>mango</td>
<td>chair</td>
</tr>
</tbody>
</table>
Hyponymy more formally

• Extensional:
  • The class denoted by the superordinate extensionally includes the class denoted by the hyponym

• Entailment:
  • A sense A is a hyponym of sense B if being an A entails being a B

• Hyponymy is usually transitive
  • (A hypo B and B hypo C entails A hypo C)

• Another name: the IS-A hierarchy
  • A IS-A B (or A ISA B)
  • B subsumes A
Hyponyms and Instances

- WordNet has both **classes** and **instances**.
- An **instance** is an individual, a proper noun that is a unique entity
  - San Francisco is an **instance** of city
- But **city** is a class
  - **city** is a **hyponym** of municipality...location...
Word Meaning and Similarity

WordNet and other Online Thesauri
Applications of Thesauri and Ontologies

- Information Extraction
- Information Retrieval
- Question Answering
- Bioinformatics and Medical Informatics
- Machine Translation
WordNet 3.0

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
  - Some other languages available or under development
    - (Arabic, Finnish, German, Portuguese...)

<table>
<thead>
<tr>
<th>Category</th>
<th>Unique Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>117,798</td>
</tr>
<tr>
<td>Verb</td>
<td>11,529</td>
</tr>
<tr>
<td>Adjective</td>
<td>22,479</td>
</tr>
<tr>
<td>Adverb</td>
<td>4,481</td>
</tr>
</tbody>
</table>
Senses of “bass” in Wordnet

Noun

- **S**: (n) **bass** (the lowest part of the musical range)
- **S**: (n) **bass**, **bass part** (the lowest part in polyphonic music)
- **S**: (n) **bass**, **basso** (an adult male singer with the lowest voice)
- **S**: (n) **sea bass**, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- **S**: (n) **freshwater bass**, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- **S**: (n) **bass**, **bass voice**, **basso** (the lowest adult male singing voice)
- **S**: (n) **bass** (the member with the lowest range of a family of musical instruments)
- **S**: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- **S**: (adj) **bass**, **deep** (having or denoting a low vocal or instrumental range) "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"
How is “sense” defined in WordNet?

- The **synset (synonym set)**, the set of near-synonyms, instantiates a sense or concept, with a **gloss**
- Example: chump as a noun with the gloss:
  “a person who is gullible and easy to take advantage of”
- This sense of “chump” is shared by 9 words:
  chump¹, fool², gull¹, mark⁹, patsy¹, fall guy¹, sucker¹, soft touch¹, mug²
- Each of **these** senses have this same gloss
  *(Not every sense; sense 2 of gull is the aquatic bird)*
WordNet Hypernym Hierarchy for “bass”

- **S:** (n) bass, *basso* (an adult male singer with the lowest voice)
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - **S:** (n) singer, vocalist, vocalizer, vocaliser (a person who sings)
    - **S:** (n) musician, instrumentalist, player (someone who plays a musical instrument (as a profession))
    - **S:** (n) performer, performing artist (an entertainer who performs a dramatic or musical work for an audience)
    - **S:** (n) entertainer (a person who tries to please or amuse)
  - **S:** (n) person, individual, someone, somebody, mortal, soul (a human being) "there was too much for one person to do"
  - **S:** (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
    - **S:** (n) living thing, animate thing (a living (or once living) entity)
    - **S:** (n) whole, unit (an assemblage of parts that is regarded as a single entity) "how big is that part compared to the whole?"; "the team is a unit"
  - **S:** (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) "it was full of rackets, balls and other objects"
    - **S:** (n) physical entity (an entity that has physical existence)
    - **S:** (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
# WordNet Noun Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Also called</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym</td>
<td>Superordinate</td>
<td>From concepts to superordinates</td>
<td>breakfast(^1) → meal(^1)</td>
</tr>
<tr>
<td></td>
<td>Subordinate</td>
<td>From concepts to subtypes</td>
<td>meal(^1) → lunch(^1)</td>
</tr>
<tr>
<td></td>
<td>Has-Member</td>
<td>From groups to their members</td>
<td>faculty(^2) → professor(^1)</td>
</tr>
<tr>
<td>Member Meronym</td>
<td>Member-Of</td>
<td>From concepts to instances of the concept</td>
<td>composer(^1) → Bach(^1)</td>
</tr>
<tr>
<td>Has-Instance</td>
<td>Has-Part</td>
<td>From instances to their concepts</td>
<td>Austen(^1) → author(^1)</td>
</tr>
<tr>
<td>Instance</td>
<td>Part-Of</td>
<td>From members to their groups</td>
<td>copilot(^1) → crew(^1)</td>
</tr>
<tr>
<td>Member Holonym</td>
<td></td>
<td>From wholes to parts</td>
<td>table(^2) → leg(^3)</td>
</tr>
<tr>
<td>Part Meronym</td>
<td></td>
<td>From parts to wholes</td>
<td>course(^7) → meal(^1)</td>
</tr>
<tr>
<td>Part Holonym</td>
<td></td>
<td>Opposites</td>
<td>leader(^1) → follower(^1)</td>
</tr>
<tr>
<td>Antonym</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MeSH: Medical Subject Headings
thesaurus from the National Library of Medicine

• MeSH (Medical Subject Headings)
  • 177,000 entry terms that correspond to 26,142 biomedical “headings”

• Hemoglobins
  
  **Entry Terms:** Eryhem, Ferrous Hemoglobin, Hemoglobin
  **Definition:** The oxygen-carrying proteins of ERYTHROCYTES. They are found in all vertebrates and some invertebrates. The number of globin subunits in the hemoglobin quaternary structure differs between species. Structures range from monomeric to a variety of multimeric arrangements
The MeSH Hierarchy

Amino Acids, Peptides, and Proteins [D12]
Proteins [D12.776]

Blood Proteins [D12.776.124]

Acute-Phase Proteins [D12.776.124.050] +

Anion Exchange Protein 1, Erythrocyte [D12.776.124.078]

Ankyrins [D12.776.124.080]

beta 2-Glycoprotein I [D12.776.124.117]

Blood Coagulation Factors [D12.776.124.125] +

Cholesterol Ester Transfer Proteins [D12.776.124.197]

Fibrin [D12.776.124.270] +

Glycophorin [D12.776.124.300]

Hemocyanin [D12.776.124.337]

Hemoglobins [D12.776.124.400]

Carboxyhemoglobin [D12.776.124.400.141]

Erythrocruiorins [D12.776.124.400.220]
Uses of the MeSH Ontology

- Provide synonyms ("entry terms")
  - E.g., glucose and dextrose
- Provide hypernyms (from the hierarchy)
  - E.g., glucose ISA monosaccharide
- Indexing in MEDLINE/PubMED database
  - NLM’s bibliographic database:
    - 20 million journal articles
    - Each article hand-assigned 10-20 MeSH terms
Word Meaning and Similarity

Word Similarity: Thesaurus Methods
**Word Similarity**

- **Synonymy**: a binary relation
  - Two words are either synonymous or not
- **Similarity (or distance)**: a looser metric
  - Two words are more similar if they share more features of meaning
- Similarity is properly a relation between **senses**
  - The word “bank” is not similar to the word “slope”
    - Bank$^1$ is similar to fund$^3$
    - Bank$^2$ is similar to slope$^5$
- But we’ll compute similarity over both words and senses
Why word similarity

- Information retrieval
- Question answering
- Machine translation
- Natural language generation
- Language modeling
- Automatic essay grading
- Plagiarism detection
- Document clustering
Word similarity and word relatedness

• We often distinguish **word similarity** from **word relatedness**
  • Similar words: near-synonyms
  • Related words: can be related any way
    • car, bicycle: **similar**
    • car, gasoline: **related**, not similar
Two classes of similarity algorithms

• Thesaurus-based algorithms
  • Are words “nearby” in hypernym hierarchy?
  • Do words have similar glosses (definitions)?

• Distributional algorithms
  • Do words have similar distributional contexts?
Path based similarity

- Two concepts (senses/synsets) are similar if they are near each other in the thesaurus hierarchy
  - have a short path between them
  - concepts have path 1 to themselves
Refinements to path-based similarity

- $\text{pathlen}(c_1, c_2) = 1 + \text{number of edges in the shortest path in the hypernym graph between sense nodes } c_1 \text{ and } c_2$
- ranges from 0 to 1 (identity)

- $\text{simpath}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)}$

- $\text{wordsim}(w_1, w_2) = \max_{c_1 \in \text{senses}(w_1), c_2 \in \text{senses}(w_2)} \text{sim}(c_1, c_2)$
Problem with basic path-based similarity

• Assumes each link represents a uniform distance
  • But *nickel* to *money* seems to us to be closer than *nickel* to *standard*
    • Nodes high in the hierarchy are very abstract
• We instead want a metric that
  • Represents the cost of each edge independently
  • Words connected only through abstract nodes
    • are less similar
Information content similarity metrics

Resnik 1995. Using information content to evaluate semantic similarity in a taxonomy. IJCAI

- Let’s define $P(c)$ as:
  - The probability that a randomly selected word in a corpus is an instance of concept $c$
  - Formally: there is a distinct random variable, ranging over words, associated with each concept in the hierarchy
    - for a given concept, each observed noun is either
      - a member of that concept with probability $P(c)$
      - not a member of that concept with probability $1-P(c)$
  - All words are members of the root node (Entity)
    - $P(\text{root})=1$
  - The lower a node in hierarchy, the lower its probability
Information content similarity

Train by counting in a corpus

- Each instance of hill counts toward frequency of natural elevation, geological formation, entity, etc.
- Let words(c) be the set of all words that are children of node c
  - words("geo-formation") = {hill, ridge, grotto, coast, cave, shore, natural elevation}
  - words("natural elevation") = {hill, ridge}

\[
P(c) = \frac{\sum_{w \in \text{words}(c)} \text{count}(w)}{N}
\]

N: vocabulary
Information content: definitions

- Information content:
  \[ IC(c) = -\log P(c) \]

- Lowest common subsumer
  \[ LCS(c_1, c_2) = \]
  The lowest node in the hierarchy that subsumes both \( c_1 \) and \( c_2 \)

- How to use information content IC as a similarity metric?
Resnik method

- The similarity between two words is related to their common information
- The more two words have in common, the more similar they are
- Resnik: measure common information as:
  - The information content of the lowest common subsumer of the two nodes
  - \( \text{sim}_{\text{resnik}}(c_1,c_2) = -\log P(\text{LCS}(c_1,c_2)) \)
The (extended) Lesk Algorithm

- A thesaurus-based measure that looks at glosses
- Two concepts are similar if their glosses contain similar words
  - *Drawing paper*: paper that is specially prepared for use in drafting
  - *Decal*: the art of transferring designs from specially prepared paper to a wood or glass or metal surface

- For each $n$-word phrase that’s in both glosses
  - Add a score of $n^2$
  - *Paper* and specially prepared for $1 + 2^2 = 5$
  - Compute overlap also for other relations
    - glosses of hypernyms and hyponyms
Summary: thesaurus-based similarity

$$\text{sim}_{\text{path}}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)}$$

$$\text{sim}_{\text{resnik}}(c_1, c_2) = -\log P(\text{LCS}(c_1, c_2)) \quad \text{sim}_{\text{lin}}(c_1, c_2) = \frac{2 \log P(\text{LCS}(c_1, c_2))}{\log P(c_1) + \log P(c_2)}$$

$$\text{sim}_{\text{jiangconrath}}(c_1, c_2) = \frac{1}{\log P(c_1) + \log P(c_2) - 2 \log P(\text{LCS}(c_1, c_2))}$$

$$\text{sim}_{\text{eLesk}}(c_1, c_2) = \sum_{r,q \in \text{RELs}} \text{overlap}(\text{gloss}(r(c_1)), \text{gloss}(q(c_2)))$$
Libraries for computing thesaurus-based similarity

- **NLTK**

- **WordNet::Similarity**
  - Web-based interface:
    - [http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi](http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi)
Evaluating similarity

- **Intrinsic Evaluation:**
  - Correlation between algorithm and human word similarity ratings

- **Extrinsic (task-based, end-to-end) Evaluation:**
  - Malapropism (spelling error) detection
  - WSD
  - Essay grading
  - Taking TOEFL multiple-choice vocabulary tests

**Levied** is closest in meaning to:
  - imposed, believed, requested, correlated
Word Meaning and Similarity

Word Similarity: Distributional Similarity (I)
Problems with thesaurus-based meaning

- We don’t have a thesaurus for every language
- Even if we do, they have problems with **recall**
  - Many words are missing
  - Most (if not all) phrases are missing
  - Some connections between senses are missing
  - Thesauri work less well for verbs, adjectives
    - Adjectives and verbs have less structured hyponymy relations
Distributional models of meaning

- Also called vector-space models of meaning
- Offer much higher recall than hand-built thesauri
  - Although they tend to have lower precision
- Zellig Harris (1954): “oculist and eye-doctor ... occur in almost the same environments.... If A and B have almost identical environments we say that they are synonyms.
- Firth (1957): “You shall know a word by the company it keeps!”
Intuition of distributional word similarity

*Nida's functional equivalence in translation (focus meaning and spirit of the original text)

• Nida example:
  A bottle of **tesgüino** is on the table
  Everybody likes **tesgüino**
  **Tesgüino** makes you drunk
  We make **tesgüino** out of corn.

• From context words humans can guess **tesgüino** means
  • an alcoholic beverage like **beer**

• Intuition for algorithm:
  • Two words are similar if they have similar word contexts.
Reminder: Term-document matrix

- Each cell: count of term \( t \) in a document \( d \): \( \text{tf}_{t,d} \):
  - Each document is a count vector in \( \mathbb{N}^v \): a column below

\[
\begin{array}{cccccc}
\text{As You Like It} & \text{Twelfth Night} & \text{Julius Caesar} & \text{Henry V} \\
\text{battle} & 1 & 1 & 8 & 15 \\
\text{soldier} & 2 & 2 & 12 & 36 \\
\text{fool} & 37 & 58 & 1 & 5 \\
\text{clown} & 6 & 117 & 0 & 0 \\
\end{array}
\]
The words in a term-document matrix

- Each word is a **count vector** in $\mathbb{N}^D$: a row below

<table>
<thead>
<tr>
<th></th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>fool</td>
<td>37</td>
<td>58</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>clown</td>
<td>6</td>
<td>117</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The Term-Context matrix

- Instead of using entire documents, use smaller contexts
  - Paragraph
  - Window of 10 words
- A word is now defined by a vector over counts of context words
Sample contexts: 20 words (Brown corpus)

• equal amount of sugar, a sliced lemon, a tablespoonful of apricot preserve or jam, a pinch each of clove and nutmeg,
• on board for their enjoyment. Cautiously she sampled her first pineapple and another fruit whose taste she likened to that of

• of a recursive type well suited to programming on the digital computer. In finding the optimal R-stage policy from that of
• substantially affect commerce, for the purpose of gathering data and information necessary for the study authorized in the first section of this
Term-context matrix for word similarity

- **Two words** are similar in meaning if their context vectors are similar.

<table>
<thead>
<tr>
<th></th>
<th>aardvark</th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>pineapple</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>information</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Should we use raw counts?

- For the term-document matrix
  - We used \texttt{tf-idf} instead of raw term counts
- For the term-context matrix
  - \texttt{Positive Pointwise Mutual Information (PPMI)} is common
Pointwise Mutual Information

- **Pointwise mutual information:**
  - Do events $x$ and $y$ co-occur more than if they were independent?
  \[
  \text{PMI}(X, Y) = \log_2 \frac{P(x, y)}{P(x)P(y)}
  \]

- **PMI between two words:** (Church & Hanks 1989)
  - Do words $x$ and $y$ co-occur more than if they were independent?
  \[
  \text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \frac{P(\text{word}_1, \text{word}_2)}{P(\text{word}_1)P(\text{word}_2)}
  \]

- **Positive PMI between two words** (Niwa & Nitta 1994)
  - Replace all PMI values less than 0 with zero
Computing PPMI on a term-context matrix

- Matrix $F$ with $W$ rows (words) and $C$ columns (contexts)
- $f_{ij}$ is # of times $w_i$ occurs in context $c_j$

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}$$

$$p_{i*} = \frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}$$

$$p_{*j} = \frac{\sum_{i=1}^{W} f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}$$

$$pmi_{ij} = \log \frac{p_{ij}}{p_{i*}p_{*j}}$$

$$ppmi_{ij} = \begin{cases} 
  pmi_{ij} & \text{if } pmi_{ij} > 0 \\
  0 & \text{otherwise}
\end{cases}$$
\[
p_{ij} = \frac{f_{ij}}{\sum_{j=1}^{C} \sum_{i=1}^{W} f_{ij}}
\]

- apricot
- pineapple
- digital
- information

\[
p(w=\text{information}, c=\text{data}) = \frac{6}{19} = .32
\]

\[
p(w=\text{information}) = \frac{11}{19} = .58
\]

\[
p(c=\text{data}) = \frac{7}{19} = .37
\]

\[
p(w,c) = \begin{array}{cccccc}
\text{computer} & \text{data} & \text{pinch} & \text{result} & \text{sugar} \\
0 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 \\
2 & 1 & 0 & 1 & 0 \\
1 & 6 & 0 & 4 & 0
\end{array}
\]

\[
p(w_i) = \frac{\sum_{j=1}^{C} f_{ij}}{N}
\]

\[
p(c_j) = \frac{\sum_{i=1}^{W} f_{ij}}{N}
\]
\[ p_{mi} = \log \frac{p_{ij}}{p_i p_j} \]

\[ p(w, \text{context}) \]

<table>
<thead>
<tr>
<th></th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>pineapple</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>digital</td>
<td>0.11</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>information</td>
<td>0.05</td>
<td>0.32</td>
<td>0.00</td>
<td>0.21</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\[ p(\text{context}) \]

|        | 0.16 | 0.37 | 0.11 | 0.26 | 0.11 |

\[ p(w) \]

|        | 0.11 | 0.11 | 0.21 | 0.58 |

**pmi(information, data) = \log \left( \frac{0.32}{(0.37 \times 0.58)} \right) = \log(1.49) = 0.39**

<table>
<thead>
<tr>
<th></th>
<th>PPMI(w, context)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>computer</td>
</tr>
<tr>
<td>apricot</td>
<td>-</td>
</tr>
<tr>
<td>pineapple</td>
<td>-</td>
</tr>
<tr>
<td>digital</td>
<td>1.15</td>
</tr>
<tr>
<td>information</td>
<td>0.00</td>
</tr>
</tbody>
</table>
### Add-1 Smoothed Count\(w,\text{context}\)

<table>
<thead>
<tr>
<th></th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>pineapple</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>digital</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>information</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

### \(p(w,\text{context})\) [add-1]

<table>
<thead>
<tr>
<th></th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>pineapple</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>digital</td>
<td>0.08</td>
<td>0.05</td>
<td>0.03</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>information</td>
<td>0.05</td>
<td>0.18</td>
<td>0.03</td>
<td>0.13</td>
<td>0.03</td>
</tr>
</tbody>
</table>

\[p(w)\]

<table>
<thead>
<tr>
<th></th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0.18</td>
<td>0.28</td>
<td>0.15</td>
<td>0.23</td>
<td>0.15</td>
</tr>
<tr>
<td>pineapple</td>
<td>0.18</td>
<td>0.28</td>
<td>0.15</td>
<td>0.23</td>
<td>0.15</td>
</tr>
<tr>
<td>digital</td>
<td>0.18</td>
<td>0.28</td>
<td>0.15</td>
<td>0.23</td>
<td>0.15</td>
</tr>
<tr>
<td>information</td>
<td>0.18</td>
<td>0.28</td>
<td>0.15</td>
<td>0.23</td>
<td>0.15</td>
</tr>
</tbody>
</table>

\[p(\text{context})\]
Using syntax to define a word’s context

- Zellig Harris (1968)
  - “The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities”
- Two words are similar if they have similar parse contexts
- **Duty** and **responsibility** (Chris Callison-Burch’s example)

<table>
<thead>
<tr>
<th>Modified by adjectives</th>
<th>additional, administrative, assumed, collective, congressional, constitutional ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objects of verbs</td>
<td>assert, assign, assume, attend to, avoid, become, breach ...</td>
</tr>
</tbody>
</table>
Co-occurrence vectors based on syntactic dependencies

Dekang Lin, 1998 “Automatic Retrieval and Clustering of Similar Words”

- The contexts C are different dependency relations
  - Subject-of- “absorb”
  - Prepositional-object of “inside”
- Counts for the word cell:

<table>
<thead>
<tr>
<th></th>
<th>subj-of, absorb</th>
<th>subj-of, adapt</th>
<th>subj-of, behave</th>
<th>pobj-of, inside</th>
<th>pobj-of, into</th>
<th>nmod-of, abnormality</th>
<th>nmod-of, anemia</th>
<th>nmod-of, architecture</th>
<th>obj-of, attack</th>
<th>obj-of, call</th>
<th>obj-of, come from</th>
<th>obj-of, decor</th>
<th>nmod, bacteria</th>
<th>nmod, body</th>
<th>nmod, bone marrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>cell</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>16</td>
<td>30</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>6</td>
<td>11</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
PMI applied to dependency relations

Hindle, Don. 1990. Noun Classification from Predicate-Argument Structure. ACL

<table>
<thead>
<tr>
<th>Object of “drink”</th>
<th>Count</th>
<th>PMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>tea</td>
<td>2</td>
<td>11.8</td>
</tr>
<tr>
<td>liquid</td>
<td>2</td>
<td>10.5</td>
</tr>
<tr>
<td>wine</td>
<td>2</td>
<td>9.3</td>
</tr>
<tr>
<td>anything</td>
<td>3</td>
<td>5.2</td>
</tr>
<tr>
<td>it</td>
<td>3</td>
<td>1.3</td>
</tr>
</tbody>
</table>

- “Drink it” more common than “drink wine”
- But “wine” is a better “drinkable” thing than “it”
Reminder: cosine for computing similarity

Cosine similarity is defined as:

$$\cos(v, w) = \frac{\vec{v} \cdot \vec{w}}{||\vec{v}|| \cdot ||\vec{w}||} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \cdot \sqrt{\sum_{i=1}^{N} w_i^2}}$$

$v_i$ is the PPMI value for word $v$ in context $i$

$w_i$ is the PPMI value for word $w$ in context $i$.

$\cos(v, w)$ is the cosine similarity of $v$ and $w$.
Cosine as a similarity metric

-1: vectors point in opposite directions
+1: vectors point in same directions
0: vectors are orthogonal

- Raw frequency or PPMI are non-negative, so cosine range 0-1
Other possible similarity measures

\[
\text{sim}_{\text{cosine}}(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}| |\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}}
\]

\[
\text{sim}_{\text{Jaccard}}(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^{N} \min(v_i, w_i)}{\sum_{i=1}^{N} \max(v_i, w_i)}
\]

\[
\text{sim}_{\text{Dice}}(\vec{v}, \vec{w}) = \frac{2 \times \sum_{i=1}^{N} \min(v_i, w_i)}{\sum_{i=1}^{N} (v_i + w_i)}
\]

\[
\text{sim}_{\text{JS}}(\vec{v} | \vec{w}) = D(\vec{v} | \frac{\vec{v} + \vec{w}}{2}) + D(\vec{w} | \frac{\vec{v} + \vec{w}}{2})
\]

JS: Jensen-Shannon
What is Sentiment Analysis?
Positive or negative movie review?

- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.
Google Product Search

HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner
$89 online, $100 nearby ★★★☆☆ 377 reviews
September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sheets

Reviews

Summary - Based on 377 reviews

<table>
<thead>
<tr>
<th>1 star</th>
<th>2</th>
<th>3</th>
<th>4 stars</th>
<th>5 stars</th>
</tr>
</thead>
</table>

What people are saying

- **ease of use**: "This was very easy to setup to four computers."
- **value**: "Appreciate good quality at a fair price."
- **setup**: "Overall pretty easy setup."
- **customer service**: "I DO like honest tech support people."
- **size**: "Pretty Paper weight."
- **mode**: "Photos were fair on the high quality mode."
- **colors**: "Full color prints came out with great quality."

110
Bing Shopping

HP Officejet 6500A E710N Multifunction Printer

Product summary  Find best price  Customer reviews  Specifications  Related items

$121.53 - $242.39 (14 stores)

Average rating ★★★★★ (144)

Most mentioned

- Performance (55)
- Ease of Use (54)
- Print Speed (14)
- Connectivity (10)
- More ▼

Show reviews by source

- Best Buy (140)
- CNET (5)
- Amazon.com (3)
Twitter sentiment versus Gallup Poll of Consumer Confidence


window = 15, r = 0.804

Sept. 15, 2008: Lehman collapse, AIG bailout
Feb 2009: Stock market bottoms out, begins recovery

Gallup Poll
Twitter Sentiment
Target Sentiment on Twitter

Type in a word and we'll highlight the good and the bad

Twitter Sentiment App

"united airlines" Search Save this search

Sentiment analysis for "united airlines"

Sentiment by Percent

Negative (68%) Positive (32%)

Sentiment by Count

Positive (11) Negative (23)

jjacobson: OMG... Could @United airlines have worse customer service? W8g now 15 minutes on hold 4 questions about a flight 2DAY that need a human.
Posted 2 hours ago

12345clumsy6789: I hate United Airlines Ceiling!!! Fukn impossible to get my conduit in this damn mess! ?
Posted 2 hours ago

EMLandPRGbelguiu: EML/PRG fly with Q8 united airlines and 24seven to an exotic destination. http://t.co/Z9QloAjF
Posted 2 hours ago

CountAdam: FANTASTIC customer service from United Airlines at XNA today. Is tweet more, but cell phones off now!
Posted 4 hours ago
Sentiment analysis has many other names

• Opinion extraction
• Opinion mining
• Sentiment mining
• Subjectivity analysis
Why sentiment analysis?

- **Movie**: is this review positive or negative?
- **Products**: what do people think about the new iPhone?
- **Public sentiment**: how is consumer confidence? Is despair increasing?
- **Politics**: what do people think about this candidate or issue?
- **Prediction**: predict election outcomes or market trends from sentiment
Scherer Typology of Affective States

- **Emotion**: brief organically synchronized ... evaluation of a major event
  - angry, sad, joyful, fearful, ashamed, proud, elated
- **Mood**: diffuse non-caused low-intensity long-duration change in subjective feeling
  - cheerful, gloomy, irritable, listless, depressed, buoyant
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - friendly, flirtatious, distant, cold, warm, supportive, contemptuous
- **Attitudes**: enduring, affectively colored beliefs, dispositions towards objects or persons
  - liking, loving, hating, valuing, desiring
- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - nervous, anxious, reckless, morose, hostile, jealous
Sentiment Analysis

- Sentiment analysis is the detection of **attitudes**
  “enduring, affectively colored beliefs, dispositions towards objects or persons”

1. **Holder (source)** of attitude
2. **Target (aspect)** of attitude
3. **Type** of attitude
   - From a set of types
     - *Like, love, hate, value, desire,* etc.
   - Or (more commonly) simple weighted **polarity**:
     - *positive, negative, neutral,* together with *strength*
4. **Text** containing the attitude
   - Sentence or entire document
Sentiment Analysis

• Simplest task:
  • Is the attitude of this text positive or negative?

• More complex:
  • Rank the attitude of this text from 1 to 5

• Advanced:
  • Detect the target, source, or complex attitude types
Sentiment Classification in Movie Reviews


- Polarity detection:
  - Is an IMDB movie review positive or negative?

- Data: Polarity Data 2.0:
  - [http://www.cs.cornell.edu/people/pabo/movie-review-data](http://www.cs.cornell.edu/people/pabo/movie-review-data)
when _star wars_ came out some twenty years ago, the image of traveling throughout the stars has become a commonplace image. [...] when han solo goes light speed, the stars change to bright lines, going towards the viewer in lines that converge at an invisible point. cool. _october sky_ offers a much simpler image—that of a single white dot, traveling horizontally across the night sky. [. . .] “snake eyes” is the most aggravating kind of movie: the kind that shows so much potential then becomes unbelievably disappointing. it’s not just because this is a brian depalma film, and since he’s a great director and one who’s films are always greeted with at least some fanfare. and it’s not even because this was a film starring nicolas cage and since he gives a brauvara performance, this film is hardly worth his talents.
Baseline Algorithm (adapted from Pang and Lee)

- Tokenization
- Feature Extraction
- Classification using different classifiers
  - Naïve Bayes
  - MaxEnt
  - SVM
Sentiment Tokenization Issues

• Deal with HTML and XML markup
• Twitter mark-up (names, hash tags)
• Capitalization (preserve for words in all caps)
• Phone numbers, dates
• Emoticons
• Useful code:
  • [Christopher Potts sentiment tokenizer](#)
  • [Brendan O’Connor twitter tokenizer](#)
Extracting Features for Sentiment Classification

- How to handle negation
  - *I didn’t like this movie*
  - *I really like this movie*

- Which words to use?
  - Only adjectives
  - All words
    - All words turns out to work better, at least on this data
Negation

Add NOT_ to every word between negation and following punctuation:

didn’t like this movie , but I

didn’t NOT_like NOT_this NOT_movie but I
Binarized (Boolean feature) Multinomial Naïve Bayes

- Intuition:
  - For sentiment (and probably for other text classification domains)
  - Word occurrence may matter more than word frequency
    - The occurrence of the word *fantastic* tells us a lot
    - The fact that it occurs 5 times may not tell us much more.
  - Boolean Multinomial Naïve Bayes
    - Clips all the word counts in each document at 1
## Normal vs. Boolean Multinomial NB

<table>
<thead>
<tr>
<th>Normal</th>
<th>Doc</th>
<th>Words</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1</td>
<td>Chinese Beijing Chinese</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Chinese Chinese Shanghai</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Chinese Macao</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Tokyo Japan Chinese</td>
<td>j</td>
</tr>
<tr>
<td>Test</td>
<td>5</td>
<td>Chinese Chinese Chinese Tokyo Japan</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Boolean</th>
<th>Doc</th>
<th>Words</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1</td>
<td>Chinese Beijing</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Chinese Shanghai</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Chinese Macao</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Tokyo Japan Chinese</td>
<td>j</td>
</tr>
<tr>
<td>Test</td>
<td>5</td>
<td>Chinese Tokyo Japan</td>
<td>?</td>
</tr>
</tbody>
</table>
Problems:
What makes reviews hard to classify?

• Subtlety:
  • Perfume review in *Perfumes: the Guide*:
    • “If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”
  • Dorothy Parker on Katherine Hepburn
    • “She runs the gamut of emotions from A to B”
Thwarted Expectations and Ordering Effects

• “This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can’t hold up.” - neg

• Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is not so good either, I was surprised. - neg
Sentiment Analysis

Sentiment Lexicons
The General Inquirer


• Home page: [http://www.wjh.harvard.edu/~inquirer](http://www.wjh.harvard.edu/~inquirer)
• List of Categories: [http://www.wjh.harvard.edu/~inquirer/homecat.htm](http://www.wjh.harvard.edu/~inquirer/homecat.htm)
• Spreadsheet: [http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls](http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls)
• Categories:
  • Positiv (1915 words) and Negativ (2291 words)
  • Strong vs Weak, Active vs Passive, Overstated versus Understated
  • Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc
• Free for Research Use
LIWC (Linguistic Inquiry and Word Count)


- 2300 words, >70 classes
- **Affective Processes**
  - negative emotion *(bad, weird, hate, problem, tough)*
  - positive emotion *(love, nice, sweet)*
- **Cognitive Processes**
  - Tentative *(maybe, perhaps, guess)*, Inhibition *(block, constraint)*
- **Pronouns, Negation** *(no, never)*, **Quantifiers** *(few, many)*
- $30 or $90 fee
SentiWordNet

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010 SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. LREC-2010

• Home page: http://sentiwordnet.isti.cnr.it/

• All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness

• [estimable(J,3)] “may be computed or estimated”
  
  Pos 0  Neg 0  Obj 1

• [estimable(J,1)] “deserving of respect or high regard”
  
  Pos .75  Neg 0  Obj .25
## Disagreements between polarity lexicons

Christopher Potts, [*Sentiment Tutorial*], 2011

<table>
<thead>
<tr>
<th></th>
<th>Opinion Lexicon</th>
<th>General Inquirer</th>
<th>SentiWordNet</th>
<th>LIWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPQA</td>
<td>33/5402 (0.6%)</td>
<td>49/2867 (2%)</td>
<td>1127/4214 (27%)</td>
<td>12/363 (3%)</td>
</tr>
<tr>
<td>Opinion Lexicon</td>
<td>32/2411 (1%)</td>
<td>1004/3994 (25%)</td>
<td></td>
<td>9/403 (2%)</td>
</tr>
<tr>
<td>General Inquirer</td>
<td></td>
<td>520/2306 (23%)</td>
<td></td>
<td>1/204 (0.5%)</td>
</tr>
<tr>
<td>SentiWordNet</td>
<td></td>
<td></td>
<td>174/694 (25%)</td>
<td></td>
</tr>
<tr>
<td>LIWC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analyzing the polarity of each word in IMDB


• How likely is each word to appear in each sentiment class?
• Count(“bad”) in 1-star, 2-star, 3-star, etc.
• But can’t use raw counts:
  • Instead, **likelihood**:
    \[
    P(w | c) = \frac{f(w, c)}{\sum_{w \in c} f(w, c)}
    \]
  • Make them comparable between words
    • **Scaled likelihood**:
    \[
    \frac{P(w | c)}{P(w)}
    \]
Analyzing the polarity of each word in IMDB


![Graphs showing the scaled likelihood of words with positive and negative polarity](image-url)
Other sentiment feature: Logical negation


- Is logical negation (*no, not*) associated with negative sentiment?
- Potts experiment:
  - Count negation (*not, n’t, no, never*) in online reviews
  - Regress against the review rating
Potts 2011 Results:
More negation in negative sentiment

IMDB (4,073,228 tokens)  Five-star reviews (846,444 tokens)

Scaled likelihood
$P(w|c)/P(w)$

1 2 3 4 5 6 7 8 9 10

1 2 3 4 5
Semi-supervised learning of lexicons

- Use a small amount of information
  - A few labeled examples
  - A few hand-built patterns
- To bootstrap a lexicon
Hatzivassiloglou and McKeown intuition for identifying word polarity


• Adjectives conjoined by “and” have same polarity
  • Fair and legitimate, corrupt and brutal
  • *fair and brutal, *corrupt and legitimate
• Adjectives conjoined by “but” do not
  • fair but brutal
Hatzivassiloglou & McKeown 1997

Step 1

- Label **seed set** of 1336 adjectives (all >20 in 21 million word WSJ corpus)
  - 657 positive
    - adequate central clever famous intelligent remarkable reputed sensitive slender thriving...
  - 679 negative
    - contagious drunken ignorant lanky listless primitive strident troublesome unresolved unsuspecting...
Hatzivassiloglou & McKeown 1997
Step 2

- Expand seed set to conjoined adjectives

Google

"was nice and"

Nice location in Porto and the front desk staff was nice and helpful...
www.tripadvisor.com/ShowUserReviews-g189180-d206904-r12068...
Mercure Porto Centro: Nice location in Porto and the front desk staff was nice and helpful - See traveler reviews, 77 candid photos, and great deals for Porto, ...

If a girl was nice and classy, but had some vibrant purple dye in ...
answers.yahoo.com › Home › All Categories › Beauty & Style › Hair
4 answers - Sep 21
Question: Your personal opinion or what you think other people's opinions might ...
Top answer: I think she would be cool and confident like katy perry :)

nice, helpful

nice, classy
Hatzivassiloglou & McKeown 1997
Step 3

- Supervised classifier assigns “polarity similarity” to each word pair, resulting in graph:
Hatzivassiloglou & McKeown 1997
Step 4

• Clustering for partitioning the graph into two
Output polarity lexicon

• Positive
  • bold decisive \textbf{disturbing} generous good honest important large mature patient peaceful positive proud sound stimulating straightforward \textbf{strange} talented vigorous witty...

• Negative
  • ambiguous \textbf{cautious} cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor \textbf{outspoken pleasant} reckless risky selfish tedious unsupported vulnerable wasteful...
Using WordNet to learn polarity


• WordNet: online thesaurus
• Create positive ("good") and negative seed-words ("terrible")
• Find Synonyms and Antonyms
  • Positive Set: Add synonyms of positive words ("well") and antonyms of negative words
  • Negative Set: Add synonyms of negative words ("awful") and antonyms of positive words ("evil")
• Repeat, following chains of synonyms
• Filter
Summary on Learning Lexicons

- **Advantages:**
  - Can be domain-specific
  - Can be more robust (more words)

- **Intuition**
  - Start with a seed set of words (‘good’, ‘poor’)
  - Find other words that have similar polarity:
    - Using “and” and “but”
    - Using words that occur nearby in the same document
    - Using WordNet synonyms and antonyms

- Use seeds and semi-supervised learning to induce lexicons
Finding sentiment of a sentence

- Important for finding aspects or attributes
  - Target of sentiment

- The food was great but the service was awful
Finding aspect/attribute/target of sentiment


- Frequent phrases + rules
  - Find all highly frequent phrases across reviews ("fish tacos")
  - Filter by rules like “occurs right after sentiment word”
    - “...great fish tacos” means fish tacos a likely aspect

<table>
<thead>
<tr>
<th>Casino</th>
<th>casino, buffet, pool, resort, beds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children’s Barber</td>
<td>haircut, job, experience, kids</td>
</tr>
<tr>
<td>Greek Restaurant</td>
<td>food, wine, service, appetizer, lamb</td>
</tr>
<tr>
<td>Department Store</td>
<td>selection, department, sales, shop, clothing</td>
</tr>
</tbody>
</table>
Finding aspect/attribute/target of sentiment

- The aspect name may not be in the sentence
- For restaurants/hotels, aspects are well-understood
- Supervised classification
  - Hand-label a small corpus of restaurant review sentences with aspect
    - food, décor, service, value, NONE
  - Train a classifier to assign an aspect to a sentence
    - “Given this sentence, is the aspect food, décor, service, value, or NONE”
Putting it all together: Finding sentiment for aspects

Results of Blair-Goldensohn et al. method

*average star rate

Rooms  (3/5 stars, 41 comments)
   (+) The room was clean and everything worked fine – even the water pressure ...
   (+) We went because of the free room and was pleasantly pleased ...
   (-) ...the worst hotel I had ever stayed at ...

Service  (3/5 stars, 31 comments)
   (+) Upon checking out another couple was checking early due to a problem ...
   (+) Every single hotel staff member treated us great and answered every ...
   (-) The food is cold and the service gives new meaning to SLOW.

Dining (3/5 stars, 18 comments)
   (+) our favorite place to stay in biloxi.the food is great also the service ...
   (+) Offer of free buffet for joining the Play
Baseline methods assume classes have equal frequencies!

- If not balanced (common in the real world)
  - can’t use accuracies as an evaluation
  - need to use F-scores
- Severe imbalancing also can degrade classifier performance
- Two common solutions:
  1. Resampling in training
     - Random undersampling
  2. Cost-sensitive learning
     - Penalize SVM more for misclassification of the rare thing
Scherer Typology of Affective States

- **Emotion**: brief organically synchronized ... evaluation of a major event
  - angry, sad, joyful, fearful, ashamed, proud, elated
- **Mood**: diffuse non-caused low-intensity long-duration change in subjective feeling
  - cheerful, gloomy, irritable, listless, depressed, buoyant
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - friendly, flirtatious, distant, cold, warm, supportive, contemptuous
- **Attitudes**: enduring, affectively colored beliefs, dispositions towards objects or persons
  - liking, loving, hating, valuing, desiring
- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - nervous, anxious, reckless, morose, hostile, jealous
Computational work on other affective states

- **Emotion:**
  - Detecting annoyed callers to dialogue system
  - Detecting confused/frustrated versus confident students

- **Mood:**
  - Finding traumatized or depressed writers

- **Interpersonal stances:**
  - Detection of flirtation or friendliness in conversations

- **Personality traits:**
  - Detection of extroverts
Detection of Friendliness

Ranganath, Jurafsky, McFarland

• Friendly speakers use collaborative conversational style
  • Laughter
  • Less use of negative emotional words
  • More sympathy
    • That’s too bad     I’m sorry to hear that
  • More agreement
    • I think so too
  • Less hedges
    • kind of     sort of     a little ...