Statistical Properties of Text

Jaime Arguello
INLS 509: Information Retrieval
jarguell@email.unc.edu
The Basic IR Process

- Information need
- Representation
- Query
- Comparison
- Retrieved objects
- Evaluation

Full-text document representation

Indexed objects
Gerard Salton (8 March 1927 in Nuremberg - 28 August 1995), also known as Gerry Salton, was a Professor of Computer Science at Cornell University. Salton was perhaps the leading computer scientist working in the field of information retrieval during his time. His group at Cornell developed the SMART Information Retrieval System, which he initiated when he was at Harvard.

- Mark-up removal
- Down-casing
- Tokenization
Our goal is to describe content using content

Are all these words equally descriptive?

What are the most descriptive words?

How might a computer identify these?
Statistical Properties of Text

• We know that language use is varied
• There are many ways to convey the same information (which makes IR difficult)
• However, are there statistical properties of word usage that are predictable? Across languages? Across modalities? Across genres?
Each document corresponds to a movie, a plot description, and a list of artists and their roles

- number of documents: 230,721
- number of term occurrences (tokens): 36,989,629
- number of unique terms (token-types): 424,035

http://www.imdb.com/
## IMDB Corpus
### term-frequencies

<table>
<thead>
<tr>
<th>rank</th>
<th>term</th>
<th>frequency</th>
<th>rank</th>
<th>term</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>1586358</td>
<td>11</td>
<td>year</td>
<td>250151</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>854437</td>
<td>12</td>
<td>he</td>
<td>242508</td>
</tr>
<tr>
<td>3</td>
<td>and</td>
<td>822091</td>
<td>13</td>
<td>movie</td>
<td>241551</td>
</tr>
<tr>
<td>4</td>
<td>to</td>
<td>804137</td>
<td>14</td>
<td>her</td>
<td>240448</td>
</tr>
<tr>
<td>5</td>
<td>of</td>
<td>657059</td>
<td>15</td>
<td>artist</td>
<td>236286</td>
</tr>
<tr>
<td>6</td>
<td>in</td>
<td>472059</td>
<td>16</td>
<td>character</td>
<td>234754</td>
</tr>
<tr>
<td>7</td>
<td>is</td>
<td>395968</td>
<td>17</td>
<td>cast</td>
<td>234202</td>
</tr>
<tr>
<td>8</td>
<td>i</td>
<td>390282</td>
<td>18</td>
<td>plot</td>
<td>234189</td>
</tr>
<tr>
<td>9</td>
<td>his</td>
<td>328877</td>
<td>19</td>
<td>for</td>
<td>207319</td>
</tr>
<tr>
<td>10</td>
<td>with</td>
<td>253153</td>
<td>20</td>
<td>that</td>
<td>197723</td>
</tr>
</tbody>
</table>
# IMDB Corpus
term-frequencies

<table>
<thead>
<tr>
<th>rank</th>
<th>term</th>
<th>frequency</th>
<th>rank</th>
<th>term</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>on</td>
<td>180760</td>
<td>31</td>
<td>their</td>
<td>116803</td>
</tr>
<tr>
<td>22</td>
<td>as</td>
<td>150721</td>
<td>32</td>
<td>they</td>
<td>116113</td>
</tr>
<tr>
<td>23</td>
<td>by</td>
<td>138580</td>
<td>33</td>
<td>has</td>
<td>113336</td>
</tr>
<tr>
<td>24</td>
<td>himself</td>
<td>138214</td>
<td>34</td>
<td>him</td>
<td>112589</td>
</tr>
<tr>
<td>25</td>
<td>but</td>
<td>134017</td>
<td>35</td>
<td>when</td>
<td>106723</td>
</tr>
<tr>
<td>26</td>
<td>she</td>
<td>132237</td>
<td>36</td>
<td>I</td>
<td>100475</td>
</tr>
<tr>
<td>27</td>
<td>who</td>
<td>132151</td>
<td>37</td>
<td>are</td>
<td>99544</td>
</tr>
<tr>
<td>28</td>
<td>an</td>
<td>129717</td>
<td>38</td>
<td>it</td>
<td>98455</td>
</tr>
<tr>
<td>29</td>
<td>from</td>
<td>122086</td>
<td>39</td>
<td>man</td>
<td>87115</td>
</tr>
<tr>
<td>30</td>
<td>at</td>
<td>118190</td>
<td>40</td>
<td>ii</td>
<td>80583</td>
</tr>
</tbody>
</table>
IMDB Corpus
term-frequencies
George Kingsley Zipf

term frequency decreases rapidly as a function of rank!
Zipf’s Law

• Term-frequency decreases rapidly as a function of rank
• How rapidly?
• Zipf’s Law:

\[ f_t = \frac{k}{r_t} \]

• \( f_t \) = frequency (number of times term \( t \) occurs)
• \( r_t \) = frequency-based rank of term \( t \)
• \( k \) = frequency of most frequent term
Zipf’s Law

\[
\frac{1}{N} \times f_t = \frac{1}{N} \times \frac{k}{r_t}
\]

\[
P_t = \frac{c}{r_t}
\]

- \(P_t = \) proportion of the collection corresponding to term \(t\)
- \(c = \) constant
- For English \(c = 0.1\) (more or less)
- What does this mean?
Zipf’s Law

\[ P_t = \frac{c}{r_t} \quad c = 0.1 \]

- The most frequent term accounts for 10% of the text
- The second most frequent term accounts for 5%
- The third most frequent term accounts for about 3%
- Together, the top 10 account for about 30%
- Together, the top 20 account for about 36%
- Together, the top 50 account for about 45%
  - that’s nearly half the text!
- What else does Zipf’s law tell us?
Zipf’s Law

- With some crafty manipulation, it also tells us that the fraction of terms that occur $n$ times is given by:

$$\frac{1}{n(n + 1)}$$

- So, what fraction of the terms occur only once?
Zipf’s Law

• With some crafty manipulation, it also tells us that the fraction of terms that occur \( n \) times is given by:

\[
\frac{1}{n(n + 1)}
\]

• About half the terms occur only once!
• About 75% of the terms occur 3 times or less!
• About 83% of the terms occur 5 times or less!
• About 90% of the terms occur 10 times or less!
Zipf’s Law

• Note: the fraction of terms that occur \( n \) times or less is given by:

\[
\sum_{i=1}^{n} \left( \frac{1}{i(i + 1)} \right)
\]

• That is, we have to add the fraction of terms that appear 1, 2, 3, ... up to \( n \) times
Verifying Zipf’s Law

visualization

Zipf’s Law

\[ f = \frac{k}{r} \]

... still Zipf’s Law

\[ \log(f) = \log\left(\frac{k}{r}\right) \]

... still Zipf’s Law

\[ \log(f) = \log(k) - \log(r) \]

- So, if Zipf’s law holds, what would we see if we plotted \( \log(f) \) vs. \( \log(r) \)?
Verifying Zipf’s Law

visualization

Zipf’s Law

\[
f = \frac{k}{r}
\]

... still Zipf’s Law

\[
\log(f) = \log\left(\frac{k}{r}\right)
\]

... still Zipf’s Law

\[
\log(f) = \log(k) - \log(r)
\]

• If Zipf’s law holds true, we should be able to plot \(\log(f)\) vs. \(\log(r)\) and see a straight line with a slope of -1
Zipf’s Law
IMDB Corpus
Does Zipf’s Law generalize across languages?
Zipf’s Law
European Parliament: English

- Transcribed speech from proceedings of the European Parliament (Koehn ’05)
Zipf’s Law

European Parliament: Spanish
Zipf’s Law
European Parliament: Italian
Zipf’s Law
European Parliament: Portuguese
Zipf’s Law
European Parliament: German
Zipf’s Law
European Parliament: Finnish
Zipf’s Law
European Parliament: Hungarian
Yes, but these texts are translations of the same content!

What about different texts?
  different topics?
  different genres?
  different sizes?
  different complexity?
Zipf’s Law
Alice in Wonderland

(text courtesy of Project Gutenberg)
Zipf’s Law
Peter Pan

(text courtesy of Project Gutenberg)
Zipf’s Law
Moby Dick

(text courtesy of Project Gutenberg)
Zipf’s Law
War and Peace

(text courtesy of Project Gutenberg)
Zipf’s Law
On the Origin of Species

(text courtesy of Project Gutenberg)
Zipf’s Law
Relativity: The Special and General Theory

(text courtesy of Project Gutenberg)
Zipf’s Law
The King James Bible

(text courtesy of Project Gutenberg)
Zipf’s Law

- Zipf’s Law holds true for:
  - different sizes of text
  - different genres
  - different topics
  - different complexity of content
  - different languages
Implications of Zipf’s Law (1)

Gérard Salton 8 March 1978 in Nuremberg 28 August 1995 also known as Gerry Salton was professor of computer science at Cornell University Salton was perhaps the leading computer scientist working in the field of information retrieval during his time his group at Cornell developed the smart information retrieval system which he initiated when he was at Harvard.

- The most descriptive words are those that do not appear in every document.
- Most retrieval models exploit this idea.
- Zipf’s law allows us to automatically identify these non-descriptive terms and treat them differently.
Implications of Zipf’s Law (2)

- Ignoring the most frequent terms greatly reduces the size of the index
- The top 50 accounts for about 45% of the collection
- These have very long inverted lists
- **Warning**: these words can be important in combination with others (e.g., in proximity operators)
- **Example queries**: “to be or not to be”, “the who”, “state of the union”, “it had to be you”
Implications of Zipf’s Law (3)

• Ignoring the most frequent terms can improve retrieval efficiency (response time)

• The most frequent terms have long inverted lists

• **Alternative**: leave them in the index and remove them from the query, unless they occur in a proximity operator
Implications of Zipf’s Law (4)

- Ignoring the most frequent terms can improve retrieval effectiveness

- Very frequent terms may not be related to the main content of the doc, but may be a “quirk” of the corpus

<table>
<thead>
<tr>
<th>rank</th>
<th>term</th>
<th>frequency</th>
<th>rank</th>
<th>term</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>1586358</td>
<td>11</td>
<td>year</td>
<td>250151</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>854437</td>
<td>12</td>
<td>he</td>
<td>242508</td>
</tr>
<tr>
<td>3</td>
<td>and</td>
<td>822091</td>
<td>13</td>
<td>movie</td>
<td>241551</td>
</tr>
<tr>
<td>4</td>
<td>to</td>
<td>804137</td>
<td>14</td>
<td>her</td>
<td>240448</td>
</tr>
<tr>
<td>5</td>
<td>of</td>
<td>657059</td>
<td>15</td>
<td>artist</td>
<td>236286</td>
</tr>
<tr>
<td>6</td>
<td>in</td>
<td>472059</td>
<td>16</td>
<td>character</td>
<td>234754</td>
</tr>
<tr>
<td>7</td>
<td>is</td>
<td>395968</td>
<td>17</td>
<td>cast</td>
<td>234202</td>
</tr>
<tr>
<td>8</td>
<td>i</td>
<td>390282</td>
<td>18</td>
<td>plot</td>
<td>234189</td>
</tr>
<tr>
<td>9</td>
<td>his</td>
<td>328877</td>
<td>19</td>
<td>for</td>
<td>207319</td>
</tr>
<tr>
<td>10</td>
<td>with</td>
<td>253153</td>
<td>20</td>
<td>that</td>
<td>197723</td>
</tr>
</tbody>
</table>
We’ve talked about Zipf’s Law in the collection.

What about Zipf’s Law in queries issued to the search engine?
Implications of Zipf’s Law (5)
Implications of Zipf’s Law (5)

• **Same trend:** a few queries occur very frequently, while most occur very infrequently

• **Opportunity:** the system can be tweaked to do well on those queries it is likely to “see” again and again

• **Curse:** this is only a **partial** solution.

• In Web search, about half the queries ever observed are unique

• How does this affect evaluation?
Implications of Zipf’s Law

- Given Zipf’s Law, as a collection grows, how will the size of the vocabulary grow?
Vocabulary Growth and Heaps’ Law

• The number of new words decreases as the size of the corpus increases

• Heaps’ Law:

\[ v = k \times n^B \]

• \( v \) = size of the vocabulary (number of unique words)
• \( n \) = size of the corpus (number of word-occurrences)
• \( k \) = constant (10 ≤ \( k \) ≤ 100)
  ‣ not the same as \( k \) in Zipf’s law
• \( B \) = constant (\( B \approx 0.50 \))
Heaps’ Law
IMDB Corpus
Heaps’ Law

- As the corpus grows, the number of new terms increases dramatically at first, but then increases at a slower rate.
- Nevertheless, as the corpus grows, new terms will always be found (even if the corpus becomes huge):
  - there is no end to vocabulary growth
  - invented words, proper nouns (people, products), misspellings, email addresses, etc.
Implications of Heaps’ Law

- Given a corpus and a new set of data, the number of new index terms will depend on the size of the corpus.
- Given more data, new index terms will always be required.
- This may also be true for controlled vocabularies (?)
  - Given a corpus and a new set of data, the requirement for new concepts will depend on the size of the corpus.
  - Given more data, new concepts will always be required.
Term Co-occurrence

- So far, we’ve talked about statistics for **single** terms
- What about statistics for **pairs** of terms?
- Term co-occurrence considers the extent to which different terms tend to appear **together** in text
- Does knowledge that one term appears, tell us whether another term is likely to appear?
Term Co-occurrence Example

war vs. peace

(Graph from The Google Books N-gram Corpus)
Term Co-occurrence Example
chocolate vs. vanilla

(The Google Books N-gram Corpus)
In-class Exercise
Word Co-occurrence

<table>
<thead>
<tr>
<th>word $w_1$ appears</th>
<th>word $w_1$ does not appear</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>b</td>
</tr>
<tr>
<td>word $w_2$ appears</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>d</td>
</tr>
</tbody>
</table>

total # of documents
$N = a + b + c + d$
A Few Important Concepts in Probability Theory and Statistics

(Some material courtesy of Andrew Moore: http://www.autonlab.org/tutorials/prob.html)
Discrete Random Variable

• A is a discrete random variable if:
  ‣ A describes an event with a finite number of possible outcomes (discrete vs. continuous)
  ‣ A describes an event whose outcome has some degree of uncertainty (random vs. pre-determined)

• A is a boolean-valued random variable if it describes an event with two outcomes: TRUE or FALSE

• Can you name some examples of boolean-valued random variables?
Boolean-Valued Random Variables

Examples

- \( A = \) it will rain tomorrow
- \( A = \) the outcome of a coin-flip will be heads
- \( A = \) The US president in 2023 will be female
- \( A = \) you have the flu
- \( A = \) the word “retrieval” will occur in a document
Probabilities

- \( P(A=\text{TRUE}) \): the probability that the outcome is \text{TRUE}  
  - the probability that it will rain tomorrow  
  - the probability that the coin will show “heads”  
  - the probability that “retrieval” appears in the doc

- \( P(A=\text{FALSE}) \): the probability that the outcome is \text{FALSE}  
  - the probability that it will NOT rain tomorrow  
  - the probability that the coin will show “tails”  
  - the probability that “retrieval” does NOT appear in the doc
Estimating the Probability of an Outcome

- $P(\text{heads}=\text{TRUE})$
- $P(\text{rain tomorrow}=\text{TRUE})$
- $P(\text{alarm sound this week}=\text{TRUE})$
- $P(\text{female pres. 2023}=\text{TRUE})$
- $P(\text{you have the flu}=\text{TRUE})$
- $P(\text{"retrieval" in a document}=\text{TRUE})$
Statistical Estimation

• Use data to **estimate** the probability of an outcome
• Data = observations of previous outcomes of the event
• What is the probability that the coin will show “heads”?
• **Statistical Estimation Example:**
  
  To gather data, you flip the coin 100 times
  You observe 54 “heads” and 46 “tails”
  What would be your estimation of $P(\text{heads} = \text{TRUE})$?
Statistical Estimation

- Use data to estimate the probability of an outcome
- Data = observations of previous outcomes of the event
- What is the probability that the coin will show “heads”?
- Statistical Estimation Example:
  - To gather data, you flip the coin 100 times
  - You observe 54 “heads” and 46 “tails”
  - What would be your estimation of $P(\text{heads}=\text{TRUE})$?
• What is the probability that it will rain tomorrow?
• Statistical Estimation Example:
  ‣ To gather data, you keep a log of the past 365 days
  ‣ You observe that it rained on 93 of those days
  ‣ What would be your estimation of $P(\text{rain}=\text{TRUE})$?
Statistical Estimation

• What is the probability that “retrieval” occurs in a document?

• Statistical Estimation Example:
  ‣ To gather data, you take a sample of 1000 documents
  ‣ You observe that “retrieval” occurs in 2 of them.
  ‣ What would be your estimation of $P(\text{"retrieval" in a document}=\text{TRUE})$?

• Usually, the more data, the better the estimation!
Joint and Conditional Probability

- For simplicity, $P(A=\text{TRUE})$ is typically written as $P(A)$
- $P(A,B)$: the probability that event $A$ and event $B$ both occur
- $P(A|B)$: the probability that event $A$ occurs given prior knowledge that event $B$ occurs
Chain Rule

- \( P(A, B) = P(A|B) \times P(B) \)

- Example:
  - probability that it will rain today and tomorrow =
  - probability that it will rain today \( \times \)
  - probability that it will rain tomorrow given prior knowledge that it rained today
Independence

• Events $A$ and $B$ are independent if:

$$P(A,B) = P(A) \times P(B)$$
Independence

- Events $A$ and $B$ are independent if:

$$P(A,B) = P(A|B) \times P(B) = P(A) \times P(B)$$

- Events $A$ and $B$ are independent if the outcome of $A$ tells us nothing about the outcome of $B$ (and vice-versa).

- Can you think of examples of two events that are (in)dependent?
Independence

• Suppose $A =$ rain tomorrow and $B =$ rain today
  ▸ Are these likely to be independent?
• Suppose $A =$ rain tomorrow and $B =$ coin flip lands ‘tails’
  ▸ Are these likely to be independent?
Mutual Information

\[ MI(w_1, w_2) = \log \left( \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \right) \]

- \( P(w_1, w_2) \): probability that words \( w_1 \) and \( w_2 \) both appear in a text
- \( P(w_1) \): probability that word \( w_1 \) appears in a text, with or without \( w_2 \)
- \( P(w_2) \): probability that word \( w_2 \) appears in a text, with or without \( w_1 \)
- The definition of “a text” is up to you (e.g., a sentence, a paragraph, a document)
Mutual Information

\[ MI(w_1, w_2) = \log \left( \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \right) \]

- If \( P(w_1, w_2) = P(w_1)P(w_2) \), it means that the words are independent: knowing that one appears conveys no information that the other one appears.

- If \( P(w_1, w_2) > P(w_1)P(w_2) \), it means that the words are not independent: knowing that one appears makes it more probable that the other one appears.
**Mutual Information**

estimation (using documents as units of analysis)

<table>
<thead>
<tr>
<th></th>
<th>word ( w_1 ) appears</th>
<th>word ( w_1 ) does not appear</th>
</tr>
</thead>
<tbody>
<tr>
<td>word ( w_2 ) appears</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>word ( w_2 ) does not appear</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

\[
P(w_1, w_2) = ?
\]

\[
P(w_1) = ?
\]

\[
P(w_2) = ?
\]

every document falls under one of these quadrants

total # of documents

\[
N = a + b + c + d
\]
**Mutual Information estimation (using documents as units of analysis)**

<table>
<thead>
<tr>
<th></th>
<th>Word $w_1$ appears</th>
<th>Word $w_1$ does not appear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word $w_2$ appears</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Word $w_2$ does not appear</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

- $P(w_1, w_2) = \frac{a}{N}$
- $P(w_1) = \frac{(a + c)}{N}$
- $P(w_2) = \frac{(a + b)}{N}$

Every document falls under one of these quadrants.

Total # of documents: $N = a + b + c + d$
## Mutual Information

### IMDB Corpus

- **Word-pairs with highest mutual information (1-20)**

<table>
<thead>
<tr>
<th>w1</th>
<th>w2</th>
<th>MI</th>
<th>w1</th>
<th>w2</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>francisco</td>
<td>san</td>
<td>6.619</td>
<td>dollars</td>
<td>million</td>
<td>5.437</td>
</tr>
<tr>
<td>angeles</td>
<td>los</td>
<td>6.282</td>
<td>brooke</td>
<td>rick</td>
<td>5.405</td>
</tr>
<tr>
<td>prime</td>
<td>minister</td>
<td>5.976</td>
<td>teach</td>
<td>lesson</td>
<td>5.370</td>
</tr>
<tr>
<td>united</td>
<td>states</td>
<td>5.765</td>
<td>canada</td>
<td>canadian</td>
<td>5.338</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>5.639</td>
<td>un</td>
<td>ma</td>
<td>5.334</td>
</tr>
<tr>
<td>winning</td>
<td>award</td>
<td>5.597</td>
<td>nicole</td>
<td>roman</td>
<td>5.255</td>
</tr>
<tr>
<td>brooke</td>
<td>taylor</td>
<td>5.518</td>
<td>china</td>
<td>chinese</td>
<td>5.231</td>
</tr>
<tr>
<td>con</td>
<td>un</td>
<td>5.514</td>
<td>japan</td>
<td>japanese</td>
<td>5.204</td>
</tr>
<tr>
<td>un</td>
<td>la</td>
<td>5.512</td>
<td>belle</td>
<td>roman</td>
<td>5.202</td>
</tr>
<tr>
<td>belle</td>
<td>nicole</td>
<td>5.508</td>
<td>border</td>
<td>mexican</td>
<td>5.186</td>
</tr>
</tbody>
</table>
## Mutual Information

**IMDB Corpus**

- Word-pairs with highest mutual information (20-40)

<table>
<thead>
<tr>
<th>w1</th>
<th>w2</th>
<th>MI</th>
<th>w1</th>
<th>w2</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>belle</td>
<td>lucas</td>
<td>5.138</td>
<td>brooke</td>
<td>eric</td>
<td>4.941</td>
</tr>
<tr>
<td>nick</td>
<td>brooke</td>
<td>5.136</td>
<td>serial</td>
<td>killer</td>
<td>4.927</td>
</tr>
<tr>
<td>loved</td>
<td>ones</td>
<td>5.116</td>
<td>christmas</td>
<td>eve</td>
<td>4.911</td>
</tr>
<tr>
<td>hours</td>
<td>24</td>
<td>5.112</td>
<td>italy</td>
<td>italian</td>
<td>4.909</td>
</tr>
<tr>
<td>magazine</td>
<td>editor</td>
<td>5.103</td>
<td>un</td>
<td>I</td>
<td>4.904</td>
</tr>
<tr>
<td>e</td>
<td>fianc</td>
<td>5.088</td>
<td>photo</td>
<td>shoot</td>
<td>4.866</td>
</tr>
<tr>
<td>newspaper</td>
<td>editor</td>
<td>5.080</td>
<td>ship</td>
<td>aboard</td>
<td>4.856</td>
</tr>
<tr>
<td>donna</td>
<td>brooke</td>
<td>5.064</td>
<td>al</td>
<td>un</td>
<td>4.800</td>
</tr>
<tr>
<td>ed</td>
<td>un</td>
<td>5.038</td>
<td>plane</td>
<td>flight</td>
<td>4.792</td>
</tr>
<tr>
<td>mexican</td>
<td>mexico</td>
<td>5.025</td>
<td>nicole</td>
<td>victor</td>
<td>4.789</td>
</tr>
</tbody>
</table>
## Mutual Information

**IMDB Corpus**

- Word-pairs with highest mutual information (1-20)

<table>
<thead>
<tr>
<th>w1</th>
<th>w2</th>
<th>MI</th>
<th>w1</th>
<th>w2</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>francisco</td>
<td>san</td>
<td>6.619</td>
<td>prime</td>
<td>un</td>
<td>5.976</td>
</tr>
<tr>
<td>angeles</td>
<td>los</td>
<td>6.282</td>
<td>united</td>
<td>con</td>
<td>5.518</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>5.639</td>
<td>un</td>
<td>ma</td>
<td>5.334</td>
</tr>
<tr>
<td>winning</td>
<td>award</td>
<td>5.597</td>
<td>nicole</td>
<td>roman</td>
<td>5.255</td>
</tr>
<tr>
<td>brooke</td>
<td>taylor</td>
<td>5.518</td>
<td>china</td>
<td>chinese</td>
<td>5.231</td>
</tr>
<tr>
<td>con</td>
<td>un</td>
<td>5.514</td>
<td>japan</td>
<td>japanese</td>
<td>5.204</td>
</tr>
<tr>
<td>un</td>
<td>la</td>
<td>5.512</td>
<td>belle</td>
<td>roman</td>
<td>5.202</td>
</tr>
<tr>
<td>belle</td>
<td>nicole</td>
<td>5.508</td>
<td>border</td>
<td>mexican</td>
<td>5.186</td>
</tr>
</tbody>
</table>

**Not a perfect metric! Subject to subtleties in the collection (these are pairs of semantically unrelated Spanish words)**
Implications of Term Co-occurrence

• Potential to improve search
  ▸ word-variants co-occur: canada, canadian
  ▸ semantically-related terms co-occur: plane, flight
  ▸ phrases describe important concepts: san francisco

• Multiple paths to improvement
  ▸ document representation: conflating variants, adding related terms, indexing phrases
  ▸ information need representation: conflating variants, adding related terms, proximity operators
  ▸ search assistance and interactions: query suggestions
Implications of Term Co-occurrence (1)
Implications of Term Co-occurrence (2)
Take-Home Message

• Language use is highly varied
• However, there are statistical properties of language that are highly consistent across domains and languages
• A few terms occur very frequently and most terms occur very infrequently
• Term co-occurrences can be used to identify semantically related terms and phrases
• These statistical properties of text make search easier
• Learn them, love them, and use them to your advantage in doing automatic analysis of text
In-class Exercise

- Find two terms with the highest MI
- 600K Washington Post Articles 2012-2017
- https://sils-jasrv.ad.unc.edu:8443/info_request/mutual_information/enter_terms.jsp