Word similarity
How similar are two words?

score: \( \text{sim}(w_1, w_2) = ? \)

\( \text{rank: } W_1 \quad W_2 \quad W_3 \)

list: \( w_1 \) and \( w_2 \) are synonyms

Word similarity
Four categories of approaches (maybe more)
- Character-based
  - turned vs. truned
- Cognates (night, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts
Word similarity

Four general categories

- Character-based
  - turned vs. truned
  - cognates (night, nacht, nicht, natt, nach, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

Dictionary-based similarity

Utilize our text similarity measures

\[ \text{sim}(\text{dog}, \text{beagle}) = \text{sim}(\text{beagle}) \]

Any carnivore of the family Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.
Dictionary-based similarity

1. part of speech tagging
2. word sense disambiguation
3. most frequent sense
4. average similarity between all senses
5. max similarity between all senses
6. sum of similarity between all senses

Dictionary + WordNet

WordNet also includes a “gloss” similar to a dictionary definition

Other variants include the overlap of the word senses as well as those word senses that are related (e.g. hypernym, hyponym, etc.)

- Incorporates some of the path information as well
- Banerjee and Pedersen, 2003

Word similarity

Four general categories

- Character-based
  - turned vs. turned
  - cognates (night, nakt, nicht, nat, natt, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

Corpus-based approaches

Word  ANY blurb with the word

aardvark

eagle

dog

Ideas?
The Beagle is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter leg. Beagles are intelligent, and are popular as pets because of their size, even temper, and lack of inherited health problems.

Dogs of similar size and purpose to the modern Beagle can be traced in Ancient Greece back to around the 5th century BC. From medieval times, Beagle was used as a generic description for the smaller hounds, though these dogs differed considerably from the modern breed.

In the 1840s, a standard Beagle type was beginning to develop: the distinction between the North Country Beagle and Southern.

---

Corpus-based: feature extraction

The Beagle is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter leg.

We'd like to utilize or vector-based approach.

How could we create a vector from these occurrences?

- collect word counts from all documents with the word in it
- collect word counts from all sentences with the word in it
- collect all word counts from all words within X words of the word
- collect all words counts from words in specific relationships: subject-object, etc.

---

Word-context co-occurrence vectors

The Beagle is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter leg. Beagles are intelligent, and are popular as pets because of their size, even temper, and lack of inherited health problems.

Dogs of similar size and purpose to the modern Beagle can be traced in Ancient Greece back to around the 5th century BC. From medieval times, Beagle was used as a generic description for the smaller hounds, though these dogs differed considerably from the modern breed.

In the 1840s, a standard Beagle type was beginning to develop: the distinction between the North Country Beagle and Southern.

---

Word-context co-occurrence vectors

The Beagle is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter leg. Beagles are intelligent, and are popular as pets because of their size, even temper, and lack of inherited health problems.

Dogs of similar size and purpose to the modern Beagle can be traced in Ancient Greece back to around the 5th century BC. From medieval times, Beagle was used as a generic description for the smaller hounds, though these dogs differed considerably from the modern breed.

In the 1840s, a standard Beagle type was beginning to develop: the distinction between the North Country Beagle and Southern.

Often do some preprocessing like lowercasing and removing stop words.
Corpus-based similarity

\[ \text{sim}(\text{dog, beagle}) = \text{sim}(\text{context}_\text{vector}(\text{dog}), \text{context}_\text{vector}(\text{beagle})) \]

<table>
<thead>
<tr>
<th>the</th>
<th>5</th>
<th>the</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>its</td>
<td>1</td>
<td>its</td>
<td>1</td>
</tr>
<tr>
<td>as</td>
<td>4</td>
<td>is</td>
<td>2</td>
</tr>
<tr>
<td>breeds</td>
<td>2</td>
<td>breed</td>
<td>1</td>
</tr>
<tr>
<td>are</td>
<td>1</td>
<td>are</td>
<td>1</td>
</tr>
<tr>
<td>intelligent</td>
<td>5</td>
<td>intelligent</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Web-based similarity

Ideas?

Concatenate the snippets for the top \(N\) results

Concatenate the web page text for the top \(N\) results
Another feature weighting

TF-IDF weighting takes into account the general importance of a feature. For distributional similarity, we have the feature \( f \), but we also have the word itself \( w \) that we can use for information.

\[
\text{sim}(\text{context}_f(\text{dog}), \text{context}_f(\text{beagle}))
\]

```
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>then</td>
<td>then</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>to</td>
<td>to</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>at</td>
<td>at</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>breed</td>
<td>breed</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>are</td>
<td>are</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>intelligent</td>
<td>intelligent</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Another feature weighting

Feature weighting ideas given this additional information?

\[
\text{sim}(\text{context}_f(\text{dog}), \text{context}_f(\text{beagle}))
\]

```
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>the</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>is</td>
<td>is</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>breed</td>
<td>breed</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>are</td>
<td>are</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>intelligent</td>
<td>intelligent</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Another feature weighting

Mutual information

A bit more probability 😊

\[
I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\]

When will this be high and when will this be low?
Mutual information

A bit more probability

\[ I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \]

if \( x \) and \( y \) are independent (i.e., one occurring doesn’t impact the other occurring) then:

\[ p(x,y) = p(x)p(y) \]

What does this do to the sum?

if \( x \) and \( y \) are dependent then:

\[ p(x,y) = p(x)p(y|x) = p(y)p(x|y) \]

\[ I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{p(y)} \]

What is this asking?
When is this high?

How much more likely are we to see \( y \) given \( x \) has a particular value!
**Point-wise mutual information**

**Mutual information**

\[ I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \]

How related are two variables (i.e. over all possible values/events)

**Point-wise mutual information**

\[ PMI(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \]

How related are two particular events/values

---

**PMI weighting**

Mutual information is often used for feature selection in many problem areas

PMI weighting weights co-occurrences based on their correlation (i.e. high PMI)

**context_vector (beagle)**

- the: 2
- is: 1
- a: 2
- breed: 1
- are: 1
- intelligent: 1
- and: 1
- to: 1
- modern: 1
- ...

How do we calculate these?