PI

- http://www.youtube.com/watch?v=jG7vhMMXagQ

Class presentations

- IR (3/30)
  - Article 1: Scott and Maksym
  - Article 2: Devin and Dandre
- MT (4/11)
  - Article 1: Jonny, Chysanthia and Daniel M.
  - Article 2: Eric and Benson
- E (4/18)
  - Article 1: Kathryn and Audrey
  - Article 2: Josh and Michael
- QA (4/25)
  - Article 1: Dustin and Brennen
  - Article 2: Sam and Martin
- Summ (4/27??)
  - Article 1: Andres and Camille
  - Article 2: Jeremy and Dan F.

Admin

- Assignment 5 posted, due next Friday (4/1) at 6pm
  - can turn in by Sunday at 6pm
- Class schedule
Final project

- Read the entire handout
- Groups of 2-3 people
  - e-mail me asap if you’re looking for a group
- research-oriented project
  - must involve some evaluation
  - must be related to NLP
- Schedule
  - Monday, 4/4 project proposal
  - 4/15 status report 1
  - 4/27 status report 2
  - 5/2, 5/4 presentations
  - 5/4 writeup
- There are lots of resources out there that you can leverage

Final project ideas

- pick a text classification task
  - evaluate different machine learning methods
  - implement a machine learning method
  - analyze different feature categories
- n-gram language modeling
  - implement and compare after smoothing techniques
  - implement alternative models
- parsing
  - PCFG-based language modeling
  - lexicalized PCFG (with smoothing)
  - true n-best list generation
  - parse output reranking
  - implement another parsing approach and compare
  - parsing non-traditional domains (e.g. twitter)
- EM
  - word-alignment for text-to-text translation
  - grammar induction

Text Similarity

- A common question in NLP is how similar are texts

\[ \text{score: } \text{sim}(\text{text1}, \text{text2}) = ? \]

rank: ?
Text similarity recapped

- Set based – easy and efficient to calculate
  - word overlap
  - Jaccard
  - Dice

- Vector based
  - create a feature vector based on word occurrences (or other features)
  - Can use any distance measure
    - L1 (Manhattan)
    - L2 (Euclidean)
    - Cosine
  - Normalize the length
  - Feature/dimension weighting
    - inverse document frequency (IDF)

Stoplists: extreme weighting

- Some words like ‘a’ and ‘the’ will occur in almost every document
  - IDF will be 0 for any word that occurs in all documents
  - For words that occur in almost all the documents, they will be nearly 0

- A stoplist is a list of words that should not be considered (in this case, similarity calculations)
  - Sometimes this is the n most frequent words
  - Often, it’s a list of a few hundred words manually created

Stoplist

- Two main benefits
  - More fine grained control: some words may not be frequent, but may not have any content value (alas, teh, go)
  - Often does contain many frequent words, which can drastically reduce our storage and computation

- Any downsides to using a stoplist?
  - For some applications, some stop words may be important
Our problems

- Which of these have we addressed?
  - word order
  - length
  - synonym
  - spelling mistakes
  - word importance
  - word frequency

A model of word similarity!

Word overlap problems

A: When the defendant and his lawyer walked into the court, some of the victim supporters turned their backs to him.

B: When the defendant walked into the courthouse with his attorney, the crowd turned their backs on him.

Word similarity

- How similar are two words?

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

\[ \text{rank: } \text{W}_1 \text{ ? W}_2 \text{ W}_3 \]

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

Word similarity applications

- General text similarity
- Thesaurus generation
- Automatic evaluation
- Text-to-text
  - paraphrasing
  - summarization
  - machine translation
- Information retrieval (search)
Word similarity

- How similar are two words?

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

- rank:

\[
w \ ? \ W_1 \ W_2 \ W_3
\]

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

ideas? useful resources?

Word similarity

- Four categories of approaches (maybe more)

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

WordNet

- Lexical database for English
  - 155,287 words
  - 206,941 word senses
  - 117,659 synsets (synonym sets)
  - ~400K relations between senses
  - Parts of speech nouns, verbs, adjectives, adverbs
  - Word graph, with word senses as nodes and edges as relationships
  - Psycholinguistics
    - WN attempts to model human lexical memory
    - Design based on psychological testing
  - Created by researchers at Princeton
    - https://wordnet.princeton.edu/
  - Lots of programmatic interfaces

WordNet relations

- synonym
- antonym
- hypernyms
- hyponyms
- holonym
- meronym
- troponym
- entailment
- (and a few others)
WordNet relations

- synonym – X and Y have similar meaning
- antonym – X and Y have opposite meanings
- hypernymy – subclass
- hyponymy – superclass
- meronymy – part of
- holonymy – contains part
- troponymy – for verbs, a more specific way of doing an action
- entailment – for verbs, one activity leads to the next

WordNet: dog

**Noun**
- **dog**
- **domestic dog**
- **Canis familiaris** is a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times: occurs in many breeds: "the dog barked all night."
- **beagle** is a hypernym of dog
- **dog** is a hyponym of beagle
- **car** is a holonym of wheel
- **wheel** is a meronym of car
- **run** is a troponym of move
- **cut** is a troponym of run
- **roll** is a troponym of cut
- **(and a few others)**

**Graph, where nodes are words and edges are relationships**

There is some hierarchical information, for example with hyp-er/o-nomy.
Word similarity: Exercise

- How could you calculate word similarity if your only resource was:
  1. the words themselves
  2. WordNet
  3. a dictionary
  4. a corpus

Word similarity

- Four general categories
  - Character-based
  - Semantic web-based (e.g. WordNet)
  - Dictionary-based
  - Distributional similarity-based
    - similar words occur in similar contexts

Character-based similarity

$\text{sim}(\text{turned}, \text{truned}) = ?$

How might we do this using only the words (i.e. no outside resources?)

Edit distance (Levenshtein distance)

- The edit distance between $w_1$ and $w_2$ is the minimum number of operations to transform $w_1$ into $w_2$
- Operations:
  - insertion
  - deletion
  - substitution

$\text{EDIT}(\text{turned}, \text{truned}) = ?$
$\text{EDIT}(\text{computer}, \text{commuter}) = ?$
$\text{EDIT}(\text{banana}, \text{apple}) = ?$
$\text{EDIT}(\text{wombat}, \text{worcester}) = ?$
Edit distance

- EDIT(turned, truned) = 2
  - delete u
  - insert u
- EDIT(computer, commuter) = 1
  - replace p with m
- EDIT(banana, apple) = 5
  - delete b
  - replace a with p
  - replace n with l
  - replace a with e
- EDIT(wombat, worcester) = 6

Better edit distance

- Are all operations equally likely?
  - No
- Improvement, give different weights to different operations
  - replacing a for e is more likely than z for y
- Ideas for weightings?
  - Learn from actual data (known typos, known similar words)
  - Intuitions: phonetics
  - Intuitions: keyboard configuration

Vector character-based word similarity

\[ \text{sim}(\text{turned}, \text{truned}) = ? \]

Any way to leverage our vector-based similarity approaches from last time?

Vector character-based word similarity

\[ \text{sim}(\text{turned}, \text{truned}) = ? \]

Generate a feature vector based on the characters (or could also use the set-based measures at the character level)

problems?
Vector character-based word similarity

\[
sim(\text{restful}, \text{fluster}) = ?
\]

Character level loses a lot of information

\[
a: 0 \quad a: 0 \\
b: 0 \quad b: 0 \\
c: 0 \quad c: 0 \\
d: 1 \quad d: 1 \\
e: 1 \quad e: 1 \\
f: 0 \quad f: 0 \\
g: 0 \quad g: 0 \\
\ldots \quad \ldots
\]

Use character bigrams or even trigrams

\[
aa: 0 \quad aa: 0 \\
ab: 0 \quad ab: 0 \\
ac: 0 \quad ac: 0 \\
\ldots \quad \ldots \\
es: 1 \quad es: 1 \\
\ldots \quad \ldots \\
fs: 1 \quad fs: 1 \\
\ldots \quad \ldots \\
re: 1 \quad re: 1 \\
\ldots \quad \ldots
\]

Word similarity

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

WordNet-like Hierarchy

To utilize WordNet, we often want to think about some graph-based measure.
WordNet-like Hierarchy

- animal
  - fish
  - mammal
  - reptile
  - amphibian
  - wolf
    - horse
      - dog
        - cat
        - mare
        - stallion
        - hunting dog
        - dachshund
        - terrier

Rank the following based on similarity:
- \( \text{SIM(wolf, dog)} \)
- \( \text{SIM(wolf, amphibian)} \)
- \( \text{SIM(terrier, wolf)} \)
- \( \text{SIM(dachshund, terrier)} \)

What information/heuristics did you use to rank these?

- path length is important (but not the only thing)
- words that share the same ancestor are related
- words lower down in the hierarchy are finer grained and therefore closer

WordNet similarity measures

- path length doesn't work very well
- Some ideas:
  - path length scaled by the depth (Leacock and Chodorow, 1998)
- With a little cheating:
  - utilize the probability of a word based on the corpus frequency counts of the word and all children of that word (-log of this is the information content)
    - words higher up tend to have less information content
    - more frequent words (and ancestors of more frequent words) tend to have less information content
WordNet similarity measures

- Utilizing information content:
  - Information content of the lowest common parent (Resnik, 1995)
  - Information content of the words minus information content of the lowest common parent (Jiang and Conrath, 1997)
  - Information content of the lowest common parent divided by the information content of the words (Lin, 1998)

Word similarity

- Four general categories
  - Character-based
    - turned vs. turned
  - Cognates (night, nacht, nicht, natt, nat, noc, noch)
  - Semantic web-based (e.g. WordNet)
  - Dictionary-based
    - Distributional similarity-based
      - Similar words occur in similar contexts

Dictionary-based similarity

<table>
<thead>
<tr>
<th>Word</th>
<th>Dictionary blurb</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td>a large, nocturnal, burrowing mammal, Orycteropus afer, of central and southern Africa, feeding on ants and termites and having a long, extensile tongue, strong claws, and long ears.</td>
</tr>
<tr>
<td>beagle</td>
<td>One of a breed of small hounds having long ears, short legs, and a usually black, tan, and white coat.</td>
</tr>
<tr>
<td>dog</td>
<td>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.</td>
</tr>
</tbody>
</table>

Dictionary-based similarity

Utilize our text similarity measures

\[
\text{sim}(\text{dog}, \text{beagle}) = \frac{\text{sim}(\text{One of a breed of small hounds having long ears, short legs, and a usually black, tan, and white coat.}, \text{One of a breed of small hounds having long ears, short legs, and a usually black, tan, and white coat.})}{\text{sim}(\text{Any carnivore of the family Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.}, \text{Any carnivore of the family Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.})}
\]
Dictionary-based similarity

What about words that have multiple senses/parts of speech?

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, nacht, nach, not, nat, nac, nach)
  - Semantic web-based (e.g. WordNet)
  - Dictionary-based
    - Distributional similarity-based
      - Similar words occur in similar contexts
Corpus-based approaches

<table>
<thead>
<tr>
<th>Word</th>
<th>ANY</th>
<th>blur</th>
<th>ideas?</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beagle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
- *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

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
Beagles are intelligent, and
to the modern Beagle can be traced
From medieval times, beagle was used as
1840s, a standard Beagle type was beginning

Often do some preprocessing like lowercasing and removing stop words

Corpus-based similarity

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

Another feature weighting

- TFIDF weighting tanks 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
- This is different from traditional text similarity where we only have \( f \)
- Another feature weighting idea
  - don’t use raw co-occurrence
  - count how likely feature \( f \) and word \( w \) are to occur together
    - incorporates co-occurrence
    - but also incorporates how often \( w \) and \( f \) occur in other instances

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

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

- if \( x \) and \( y \) are independent (i.e., one occurring doesn’t impact the other occurring) \( p(x,y) = p(x)p(y) \) and the sum is 0
- if they’re dependent then \( p(x,y) = p(x)p(y|x) = p(y)p(x|y) \) then we get \( p(y|x)/p(y) \) (i.e., how much more likely are we to see \( y \) given \( x \) has a particular value) or vice versa \( p(x|y)/p(x) \)

Pointwise mutual information

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

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

PMI weighting

- Mutual information is often used for features selection in many problem areas
- PMI weighting weights co-occurrences based on their correlation (i.e., high PMI)

Web-based similarity

How can we make a document/blurb from this?

PMI (beagle, the) = \[ \log \frac{p(\text{beagle}, \text{the})}{p(\text{beagle})p(\text{the})} \]

PMI (beagle, breed) = \[ \log \frac{p(\text{beagle}, \text{breed})}{p(\text{beagle})p(\text{breed})} \]

This would likely be lower

This would likely be higher
Web-based similarity

Concatenate the snippets for the top N results

Concatenate the web page text for the top N results