TEXT SIMILARITY

David Kauchak
CS457 Fall 2011

Admin

- Assignment 4
  - Get started!

Course at a high-level

- Applications
  - Corpus analysis
  - Language modeling
  - Parsing
- For the next 3 weeks: tools
  - text similarity
  - machine learning
  - search
- Regroup for the last 3 weeks with more applications

Text Similarity

- A common question in NLP is how similar are texts

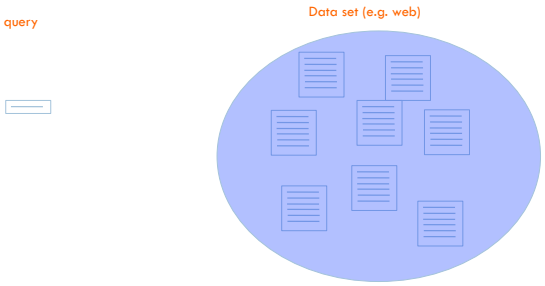
score: sim( , ) = ?

rank:  ? 

How could these be useful? Applications?

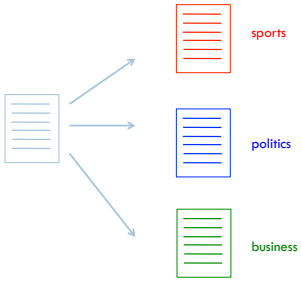
Text similarity: applications

- Information retrieval (search)



Text similarity: applications

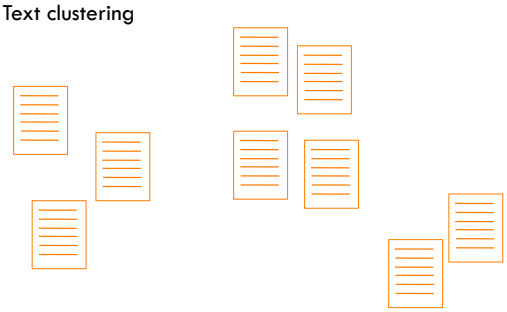
- Text classification



These "documents" could be actual documents, for example using k-means or pseudo-documents, like a class centroid/ average

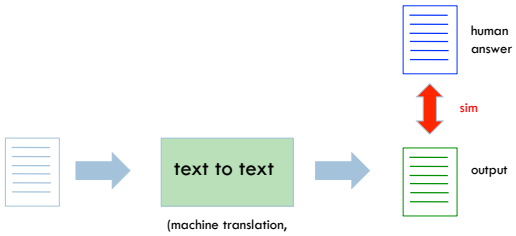
Text similarity: applications

- Text clustering



Text similarity: applications

- Automatic evaluation



(machine translation, summarization, simplification)

Text similarity: applications

- Word similarity

  sim(*banana*, apple) = ?

- Word-sense disambiguation

  I went to the *bank* to get some money.

  financial bank river bank

Text similarity: application

- Automatic grader

  Question: what is a variable?
  Answer: a location in memory that can store a value

  How good are:

  - a variable is a location in memory where a value can be stored
  - a named object that can hold a numerical or letter value
  - it is a location in the computer 's memory where it can be stored for use by a program
  - a variable is the memory address for a specific type of stored data or from a mathematical perspective a symbol representing a fixed definition with changing values
  - a location in memory where data can be stored and retrieved

Text similarity

- There are many different notions of similarity depending on the domain and the application
- Today, we'll look at some different tools
- There is no one single tool that works in all domains

Text similarity approaches

sim( , ) = ?

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 truned their backs on him.

How can we do this?

The basics: text overlap

- Texts that have overlapping words are more similar

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 truned their backs on him.

Word overlap: a numerical score

- Idea 1: number of overlapping words

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 truned their backs on him.

sim(T1, T2) = 11 problems?

Word overlap problems

- Doesn't take into account word order
- Related: doesn't reward longer overlapping sequences

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

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

sim(T1, T2) = 11

Word overlap problems

Doesn't take into account length

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 truned their backs on him. I ate a large banana at work today and thought it was great!

sim(T1, T2) = 11

Word overlap problems

Doesn't take into account synonyms

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 truned their backs on him.

sim(T1, T2) = 11

Word overlap problems

Doesn't take into account spelling mistakes

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 *truned* their backs on him.

sim(T1, T2) = 11

Word overlap problems

Treats all words the same

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 truned their backs on him.

Word overlap problems

May not handle frequency properly

A: When the defendant and his lawyer walked into the court, some of the victim supporters turned their backs to him. I ate a ***banana*** and then another ***banana*** and it was good!

B: When the defendant walked into the courthouse with his attorney, the crowd truned their backs on him. I ate a large ***banana*** at work today and thought it was great!

Word overlap: sets

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

and
backs
court
defendant
him
…

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

and
backs
courthouse
defendant
him
…

Word overlap: sets

- What is the overlap, using sets?
  - |A∧B| the size of the intersection
- How can we incorporate length/size into this measure?

Word overlap: sets

- What is the overlap, using sets?
  - |A∧B| the size of the intersection
- How can we incorporate length/size into this measure?
- Jaccard index (Jaccard similarity coefficient)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Dice's coefficient

$$Dice(A,B) = \frac{2\,|A \cap B|}{|A| + |B|}$$

Word overlap: sets

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad Dice(A,B) = \frac{2\,|A \cap B|}{|A| + |B|}$$

How are these related?

Hint: break them down in terms of

$|A - B|$ words in A but not B

$|B - A|$ words in B but not A

$|A \cap B|$ words in both A and B

Word overlap: sets

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$= \frac{|A \cap B|}{|A - B| + |B - A| + |A \cap B|}$$

in A but not B

in B but not A

$$Dice(A,B) = \frac{2\,|A \cap B|}{|A| + |B|}$$

$$= \frac{2\,|A \cap B|}{|A - B| + |B - A| + 2\,|A \cap B|}$$

Dice's coefficient gives twice the weight to overlapping words

Set overlap

- Our problems:
  - word order
  - length
  - synonym
  - spelling mistakes
  - word importance
  - word frequency

Set overlap measures can be good in some situations, but often we need more general tools

Bag of words representation

For now, let's ignore word order:

Obama said banana repeatedly last week on tv, "banana, banana, banana"

(4, 1, 1, 0, 0, 1, 0, 0, …)

banana, obama, said, california, across, tv, wrong, capital

Frequency of word occurrence

Vector based word

A

a_1: When	1
a_2: the	2
a_3: defendant	1
a_4: and	1
a_5: courthouse	0
…	

B

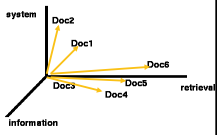
b_1: When	1
b_2: the	2
b_3: defendant	1
b_4: and	0
b_5: courthouse	1
…	

Multi-dimensional vectors, one dimension per word in our vocabulary

How do we calculate the similarity based on these vectors?

Vector based similarity

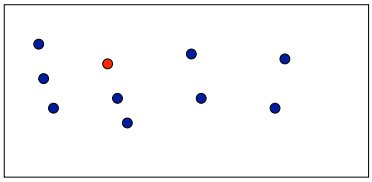
- We have a |V|-dimensional vector space
- Terms are axes of the space
- Documents are points or vectors in this space
- Very high-dimensional
- This is a very sparse vector - most entries are zero



What question are we asking in this space for similarity?

Vector based similarity

- Similarity relates to distance
- We'd like to measure the similarity of documents in the |V| dimensional space
- What are some distance measures?



Distance measures

- Euclidean (L2)

$$sim(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

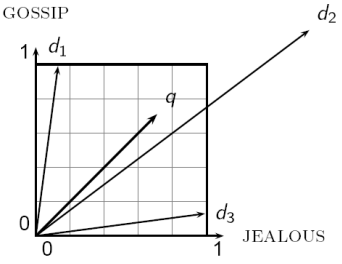
- Manhattan (L1)

$$sim(A,B) = \sum_{i=1}^{n}|a_i - b_i|$$

Distance can be problematic

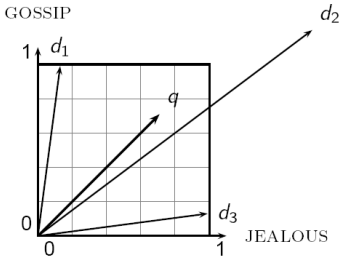
Which d is closest to q using one of the previous distance measures?

Which do you think should be closer?



Distance can be problematic

The Euclidean (or L1) distance between q and d_2 is large even though the distribution of words is similar

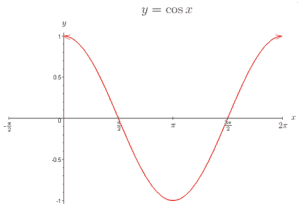


Use angle instead of distance

- Thought experiment:
  - take a document d
  - make a new document d' by concatenating two copies of d
  - "Semantically" d and d' have the same content
- What is the Euclidean distance between d and d'? What is the angle between them?
  - The Euclidean distance can be large
  - The angle between the two documents is 0

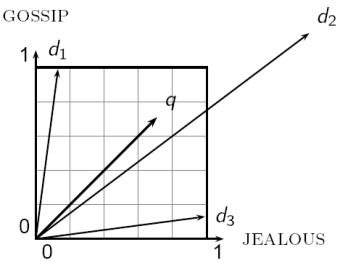
From angles to cosines

- Cosine is a monotonically decreasing function for the interval [0°, 180°]
- *decreasing* angle is equivalent to *increasing* cosine



cosine

How do we calculate the cosine between two vectors?



Cosine of two vectors

Dot product

$$A \cdot B = \|A\|\|B\|\cos\theta$$

$$\cos\theta = \frac{A \cdot B}{\|A\|\|B\|} = \frac{A}{\|A\|} \cdot \frac{B}{\|B\|}$$

Dot product between unit length vectors

Cosine as a similarity

$$sim_{\cos'}(A,B) = A \cdot B = \sum_{i=1}^{n} a_i b_i$$

ignoring length normalization

Just another distance measure, like the others:

$$sim_{L2}(A,B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

$$sim_{L1}(A,B) = \sum_{i=1}^{n} |a_i - b_i|$$

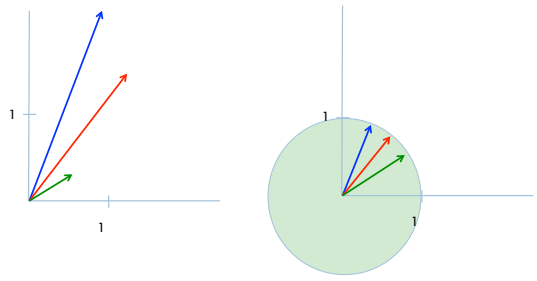
Length normalization

- A vector can be length-normalized by dividing each of its components by its length
- Often, we'll use L_2 norm (could also normalize by other norms):

$$\|\vec{x}\|_2 = \sqrt{\sum_i x_i^2}$$

- Dividing a vector by its L_2 norm makes it a unit (length) vector

Unit length vectors



In many situations, normalization improves similarity, but not in all situations

Normalized distance measures

- Cosine

$$sim_{\cos}(A,B) = A \cdot B = \sum_{i=1}^{n} a_i' b_i' = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}}$$

- L2

$$sim_{L2}(A,B) = \sqrt{\sum_{i=1}^{n} (a_i' - b_i')^2}$$

- L1

$$sim_{L1}(A,B) = \sum_{i=1}^{n} |a_i' - b_i'|$$

a' and b' are length normalized versions of the vectors