Admin

Assignment 4a
- Solutions posted
- If you’re still unsure about questions 3 and 4, come talk to me.

Assignment 4b

Grading

Quiz #2 next Thursday covering material through 10/6

Course feedback

Text Similarity

A common question in NLP is how similar are texts

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

rank: \[ ? \quad ? \quad ? \]

How could these be useful? Applications?
Text similarity: applications

Information retrieval (search)

Query
Data set (e.g. web)

Text classification

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

Text clustering

Automatic evaluation

Text to text:
(machine translation, summarization, simplification)

Human answer

Output

Sim
Text similarity: applications

Word similarity

\[ \text{sim}(\text{banana}, \text{apple}) = ? \]

Word-sense disambiguation

I went to the \textit{bank} to get some money.

\begin{itemize}
  \item financial bank
  \item river bank
\end{itemize}

Automatic grader

Question: what is a variable?

Answer: a location in memory that can store a value

How good are:

\begin{itemize}
  \item a variable is a location in memory where a value can be stored
  \item a named object that can hold a numerical or letter value
  \item it is a location in the computer's memory where it can be stored for use by a program
  \item 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
  \item a location in memory where data can be stored and retrieved
\end{itemize}

Text similarity approaches

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

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.

How can we do this?

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
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 turned 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 turned their backs on him.

\[ \text{sim}(T_1, T_2) = 11 \]

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 to him.

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

\[ \text{sim}(T_1, T_2) = 11 \]

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

\[ \text{sim}(T_1, T_2) = 11 \]
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 turned their backs on him.

\[ \text{sim}(T_1, T_2) = 11 \]

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 turned their backs on him.

\[ \text{sim}(T_1, T_2) = 11 \]

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

What is the overlap, using set notation?
- \(|A \cap B|\) 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
\[
Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|}
\]

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

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

\[ \text{Dice}(A, B) = \frac{2 |A \cap B|}{|A| + |B|} \]

\[ |A \cap B| + |A - B| + |B - A| + 2 |A \cap B| \]

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

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

For now, let’s ignore word order:

Obama said banana repeatedly last week on tv “banana, banana, banana”

What information do we lose?

Bag of words representation

(4, 1, 1, 0, 0, 1, 0, 0, …)

“Bag of words representation”: multi-dimensional vector, one dimension per word in our vocabulary

Frequency of word occurrence
Bag of words representation

10/1/20


Vector based word

A
- a: When 1
- a: the 2
- a: defendant 1
- a: and 1
- a: courthouse 0
- ...

B
- b: When 1
- b: the 2
- b: defendant 1
- b: and 0
- b: 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)
  \[ dist(A, B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \]

- Manhattan (L1)
  \[ dist(A, B) = \sum_{i=1}^{n} |a_i - b_i| \]

What do these mean for our bag of word vectors?

Distance can be problematic

Which \( d \) is closest to \( q \) using one of the previous distance measures?
Which do you think should be closer?

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 \([0^\circ, 180^\circ]\).

A decreasing angle is equivalent to an increasing cosine of that angle.
(Larger cosine means more similar.

180°: far apart
0°: close together

Near and far

https://www.youtube.com/watch?v=iZhEcRrMA-M

Cosine of two vectors

How do we calculate the cosine between two vectors?

\[
\cos \theta = \frac{A \cdot B}{\|A\| \|B\|}
\]

\[
A \cdot B = \|A\| \|B\| \cos \theta
\]

Dot product between unit length vectors
### Cosine as a similarity

\[
\text{sim}_\cos(A, B) = A \cdot B = \sum_{i=1}^{n} a_i b_i
\]

ignoring length normalization

Just another distance measure, like the others:

\[
dist_{L^2}(A, B) = \sqrt{\sum (a_i - b_i)^2}
\]

\[
dist_{L^1}(A, B) = \sum |a_i - b_i|
\]

**For bag of word vectors, what does this do?**

**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):

\[
\|v\| = \sqrt{\sum v_i^2}
\]

Dividing a vector by its L_2 norm makes it a unit (length) vector.
In many situations, normalization improves similarity, but not in all situations.

### Distance measures

**Cosine**

\[
\text{sim}_{\text{cos}}(A,B) = A \cdot B = \sum_{i=1}^{n} a_i b_i
\]

**L2**

\[
\text{dist}_{\text{L2}}(A,B) = \sqrt{\sum_{i=1}^{n} (a_i' - b_i')^2}
\]

**L1**

\[
\text{dist}_{\text{L1}}(A,B) = \sum_{i=1}^{n} |a_i' - b_i'|
\]

Cosine is the most common measure. Why do you think?

### Normalized distance measures

**Cosine**

\[
\text{sim}_{\text{cos}}(A,B) = A \cdot B = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2}}
\]

**L2**

\[
\text{dist}_{\text{L2}}(A,B) = \frac{1}{\sqrt{\sum_{i=1}^{n} (a_i' - b_i')^2}}
\]

**L1**

\[
\text{dist}_{\text{L1}}(A,B) = \sum_{i=1}^{n} |a_i' - b_i'|
\]

\[a' \text{ and } b' \text{ are length normalized versions of the vectors}\]

L1 and L2 penalize sentences for not having words, i.e., if a has it but b doesn’t.

Cosine can be significantly faster since it only calculates over the intersection.
Our problems

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

Our problems

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