Admin

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

Assignment 4b

Quiz #2 next Thursday

Admin

Office hours between now and Tuesday:
- Available Friday before 11am and 12-1pm
- Monday: 1-3pm
- Cancelled Friday and Monday original office hours

Course feedback

Thanks!

Overall, how is the class going?

1 2 3 4 5 6

0 1 2 3 4 5 6
**Course feedback**

If the exams were take-home instead of in-class (although being open book was a step in the right direction).

**Mentor sessions?**

Finish up the main lecture before the last two minutes of class. It’s hard to pay attention when I’m worrying if I’ll be able to get back to Mudd in time for Colloquium.

---

**Course feedback**

I enjoyed how the first lab had a competitive aspect to it, in comparison to the second lab which was too open ended.

I like the labs - some people seem to not get a lot out of them, but I think that it’s nice to play with actual tools, and they help reinforce concepts from lectures.

---

**Course feedback**

My favorite:

Please don’t make the next quiz too difficult. >_<
Text Similarity

A common question in NLP is how similar are texts

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

How could these be useful? Applications?

Text similarity: applications

Information retrieval (search)

query

Data set (e.g., web)

Text similarity: applications

Text classification

sports

politics

business

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

- Text to text: [machine translation, summarization, simplification]
- Human answer
- Output

Word similarity

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

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

- 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

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

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?

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 him to.

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!

Word overlap: sets

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

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

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

How are these related?

Hint: break them down in terms of

\[ |A - B| \quad \text{words in A but not B} \]
\[ |B - A| \quad \text{words in B but not A} \]
\[ |A \cap B| \quad \text{words in both A and B} \]

Dice's coefficient gives twice the weight to overlapping words

Set overlap

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

Bag of words representation

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

What information do we lose?
Bag of words representation

For now, let's ignore word order:

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

\[(4, 1, 1, 0, 1, 0, 0, \ldots)\]  

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

Frequency of word occurrence

---

Vector based word

\[
\begin{align*}
A & \quad a_1: \text{When} & 1 \\
& a_2: \text{the} & 2 \\
& a_3: \text{defendant} & 1 \\
& a_4: \text{and} & 1 \\
& a_5: \text{court} & 0 \\
& \ldots & \\
B & \quad b_1: \text{When} & 1 \\
& b_2: \text{the} & 2 \\
& b_3: \text{defendant} & 1 \\
& b_4: \text{and} & 0 \\
& b_5: \text{court} & 1 \\
& \ldots & \\
\end{align*}
\]

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)**
  \[
  \text{dist}(A,B) = \sqrt{\sum_{i} (a_i - b_i)^2}
  \]

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

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?

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

A decreasing angle is equivalent to an increasing cosine of that angle.

Near and far

https://www.youtube.com/watch?v=i2hEcRrMA-M

How do we calculate the cosine between two vectors?

\[ \cos(\theta) = \frac{d_1 \cdot d_2}{\|d_1\| \cdot \|d_2\|} \]
Cosine of two vectors:

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

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

Dot product between unit length vectors

Cosine as a similarity:

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

Just another distance measure, like the others:

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

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

Cosine as a similarity:

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

For bag of word vectors, what does this do?

Only words that occur in both documents count towards similarity

Words that occur more frequently in both receive more weight
**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):

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

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

---

**Normalized distance measures**

**Cosine**

$$\text{sim}_{\cos}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2 \sum b_i^2}}$$

**L2**

$$\text{dist}_{L2}(A, B) = \sqrt{\sum (a_i - b_i)^2}$$

**L1**

$$\text{dist}_{L1}(A, B) = \sum |a_i - b_i|$$

---

**Unit length vectors**

In many situations, normalization improves similarity, but not in all situations.
Distance measures

- **Cosine**
  \[ \text{sim}_{\text{cos}}(A, B) = A \cdot B = \sum_i a_i b_i \]

- **L2**
  \[ \text{dist}_2(A, B) = \sqrt{\sum_i (a_i - b_i)^2} \]

- **L1**
  \[ \text{dist}_1(A, B) = \sum_i |a_i - b_i| \]

- 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

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.

Ideas?
Word importance

Include a weight for each word/feature

A
- a: When 1 \( w_1 \)
- a: the 2 \( w_2 \)
- a: defendant 1 \( w_3 \)
- a: and 1 \( w_4 \)
- a: courthouse 0 \( w_5 \)
- ...

B
- b: When 1 \( w_1 \)
- b: the 2 \( w_2 \)
- b: defendant 1 \( w_3 \)
- b: and 0 \( w_4 \)
- b: courthouse 1 \( w_5 \)
- ...

Distance + weights

We can incorporate the weights into the distances

Think of it as either (both work out the same):
- preprocessing the vectors by multiplying each dimension by the weight
- incorporating it directly into the similarity measure

\[
sim_w(A, B) = \frac{A \cdot B}{\sqrt{\sum_i (w_iA_i)^2} \sqrt{\sum_i (w_iB_i)^2}}
\]

Idea: use corpus statistics

Document frequency

document frequency (DF) is one measure of word importance

Terms that occur in many documents are weighted less, since overlapping with these terms is very likely
- In the extreme case, take a word like the that occurs in almost EVERY document

Terms that occur in only a few documents are weighted more
**Document vs. overall frequency**

The overall frequency of a word is the number of occurrences in a dataset, counting multiple occurrences.

Example:

<table>
<thead>
<tr>
<th>Word</th>
<th>Overall frequency</th>
<th>Document frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>insurance</td>
<td>10440</td>
<td>3997</td>
</tr>
<tr>
<td>try</td>
<td>10422</td>
<td>8760</td>
</tr>
</tbody>
</table>

Which word is a more informative (and should get a higher weight)?

**Document frequency**

<table>
<thead>
<tr>
<th>Word</th>
<th>Collection frequency</th>
<th>Document frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>insurance</td>
<td>10440</td>
<td>3997</td>
</tr>
<tr>
<td>try</td>
<td>10422</td>
<td>8760</td>
</tr>
</tbody>
</table>

Document frequency is often related to word importance, but we want an actual weight. Problems?

**From document frequency to weight**

<table>
<thead>
<tr>
<th>Word</th>
<th>Collection frequency</th>
<th>Document frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>insurance</td>
<td>10440</td>
<td>3997</td>
</tr>
<tr>
<td>try</td>
<td>10422</td>
<td>8760</td>
</tr>
</tbody>
</table>

- weight and document frequency are **inversely** related
- Higher document frequency should have lower weight and vice versa
- document frequency is unbounded
- document frequency will change depending on the size of the data set (i.e. the number of documents)

**Inverse document frequency**

\[
\text{idf}_w = \log \frac{N}{\text{df}_w}
\]

IDF is inversely correlated with DF
- Higher DF results in lower IDF

N incorporates a dataset dependent normalizer

log dampens the overall weight
### IDF example, suppose \( N = 1 \) million

<table>
<thead>
<tr>
<th>term</th>
<th>( df_t )</th>
<th>( idf_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>calpurnia</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>animal</td>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>sunday</td>
<td>1,000</td>
<td>3</td>
</tr>
<tr>
<td>fly</td>
<td>10,000</td>
<td>2</td>
</tr>
<tr>
<td>under</td>
<td>100,000</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>1,000,000</td>
<td>0</td>
</tr>
</tbody>
</table>

**What are the IDFs assuming log base 10?**

<table>
<thead>
<tr>
<th>term</th>
<th>( df_t )</th>
<th>( idf_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>calpurnia</td>
<td>1</td>
<td>1,000,000</td>
</tr>
<tr>
<td>animal</td>
<td>100</td>
<td>10,000</td>
</tr>
<tr>
<td>sunday</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>fly</td>
<td>10,000</td>
<td>100</td>
</tr>
<tr>
<td>under</td>
<td>100,000</td>
<td>10</td>
</tr>
<tr>
<td>the</td>
<td>1,000,000</td>
<td>1</td>
</tr>
</tbody>
</table>

**What if we didn’t use the log to dampen the weighting?**

### IDF example, suppose \( N = 1 \) million

<table>
<thead>
<tr>
<th>term</th>
<th>( df_t )</th>
<th>( idf_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>calpurnia</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>animal</td>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>sunday</td>
<td>1,000</td>
<td>3</td>
</tr>
<tr>
<td>fly</td>
<td>10,000</td>
<td>2</td>
</tr>
<tr>
<td>under</td>
<td>100,000</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>1,000,000</td>
<td>0</td>
</tr>
</tbody>
</table>

**There is one \( idf \) value/weight for each word**

What if we didn’t use the log to dampen the weighting?
TF-IDF

One of the most common weighting schemes

TF = term frequency

IDF = inverse document frequency

\[ a_i' = a_i \times \log \frac{N}{df_i} \]

We can then use this with any of our similarity measures!

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 of 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

Stoplists

Two main benefits
- More fine grained control: some words may not be frequent, but may not have any content value (alas, teh, gosh)
- 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

Stoplist

If most of these end up with low weights anyway, why use a stoplist?
## Text similarity so far...

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 measures
  - L1 (Manhattan)
  - L2 (Euclidean)
  - Cosine (most common)
- Normalize the length
- Feature/dimension weighting
  - Inverse document frequency (IDF)