Alternatives to Rule-Based Approaches

Recap: Spam Classification

Social worker and able bodied . Can help bring meals to people , or simply drop off donations - can provide toiletries , blankets , jackets , hats , food , etc .

Going out to scavenge for allergy medicine . Hopefully someone is open otherwise my next painting with be splatter technique . # sandy

Went to my old job to get some food b4 this storm lmao

https://github.com/abisee/sailors2017/blob/master/data/labeled-data-singlelabels-test.csv

Are your rules perfect?

When might your rules make a mistake?

Come up with examples with your neighbors.

What are the disadvantages of rule-based approaches?

- How do you create rules that work for all of the counter-examples you came up with?
  - You need more rules, or more complicated rules
  - What happens when you have thousands and thousands of examples? How many rules might you need?

What are the disadvantages?

- How do you create rules that work for all of the counter-examples you came up with?
  - You need more rules, or more complicated rules
  - What happens when you have thousands and thousands of examples? How many rules might you need?
- What happens every time you add a new rule?
  - Go back and make sure that it works for every example before it
  - Time-consuming and difficult to make sure rules don’t contradict each other

Is there a better way?

- We can use machine learning to create “rules” automatically
- We make rules using our knowledge of language
- How can algorithms that don’t understand language make rules?

Discuss: What kind of information might be useful for an algorithm?

Similarity between article title and body

- One way to detect fake news is to find articles whose titles are misleading and not supported by their body
- What features could we look for to detect how similar an article's title is to its body?

Similarity between article title and body

- How often do the words in the title appear in the article body?
- Jaccard Similarity:

$$\frac{\text{(number of shared words)}}{\text{(number of total words)}}$$

- A higher Jaccard Similarity means more of the words overlap, so the title and the body are more likely to be similar

Jaccard Similarity

Example:

There are many possible sentences.

It is possible these sentences are similar.

What’s the Jaccard Similarity?

Jaccard Similarity

Example:

There are many possible sentences.
It is possible these sentences are similar.

What's the Jaccard Similarity?

Number of Words = 9

There, are, many, possible, sentences, it, is, these, similar

Number of shared words = 3

Are, possible, sentences

Jaccard Similarity = .33

N-Grams

- We can improve our model by not just looking at individual words
- An n-gram is *n* words together from the text
- So if we split these sentences into 2-grams (bigrams), we would get:

N-Grams

- We can improve our model by not just looking at individual words
- An n-gram is *n* words together from the text
- So if we split these sentences into 2-grams (bigrams), we would get:

  There are, are many, many possible, possible sentences

  It is, is possible, possible these, these sentences, sentences are, are similar

- What is the measure of Jaccard Similarity now?

N-Grams

Discuss:

When would a title and a body have high Jaccard similarity on bigrams (or any n-gram)?

What is this useful for?

Vector similarity

- Bag of Words representations give us vectors
- The more similar two vectors are, the more they will point in the same direction in space
- How do we calculate the angle between two vectors?

Cosine Similarity

- Use Cosine Similarity!

$$\frac{\text{(number of shared words)}}{\sqrt{\text{(number of words in title)}}\,\sqrt{\text{(number of words in body)}}}$$

- Similarity of 1 means identical and 0 means completely different

Example Cosine Similarity

Example:

There are many possible sentences.

It is possible these sentences are similar.

What is the cosine similarity?

Example Cosine Similarity

Example:

There are many possible sentences.

It is possible these sentences are similar.

Cosine Similarity: (3) / (√(5) √(7))

Cosine Similarity = 0.507