BC COMS 2710: Computational Text Analysis

Lecture 21
Phrases

Announcements

- Final Projects:
  - Presentation templates and instructions are on the final-project page on the website
  - Report/paper templates will go up later this weekend
- Course evaluations
  - Due Monday June 14th
- Office hours
  - 5-6 pm today

Announcements – HW04

- Due tonight
- Likelihoods
  - Words that don't appear in training
  - Classifying document 1
- Twitter API
  - Lecture 13 slide 2

Pre-class Instructions

1. Create a Twitter developer account https://developer.twitter.com/
2. Go to https://developer.twitter.com/en/apps and log in with your Twitter user account.
3. Click “Create an app”
4. Fill out the form, and click “Create”
5. A pop up window will appear for reviewing Developer Terms. Click the “Create” button again.

Instructions from http://socialmedia-class.org/twittertutorial.html

Today's agenda

Phrases

- n-grams
- Language models
- collocation

n-grams

N-grams

- Unigram
  - a single word
- Bigram
  - Two word phrase
- Trigram
  - Three word phrase
- 100-gram
  - One hundred word phrase
- n-gram
  - *n-word phrase*

We can add even more columns to our DTM

	w_1	w_2	w_3	w_4	$\dots$	$\dots$	$\dots$	$\dots$	w_v
d_1									
d_1									
$\dots$									
d_n									

Document-Term Matrix

We can add even more columns to our DTM

	w_1	w_2	$\dots$	$\dots$	w_v	w_1, w_2	w_1, w_3	$\dots$	w_{v-1}, w_v
d_1									
d_1									
$\dots$									
d_n									

Language Models

Probability of a word/unigram

Given a corpus C, what is the probability of a word w_i?

$$P(w_i) = \frac{count(w_i)}{\sum_j count(w_j)}$$

Maximum Likelihood Estimation

Given a corpus C, what is the probability of a word “New”?

$$P(New) = \frac{count(New)}{\sum_j count(w_j)}$$

Marginalizing

Probability of a word/unigram
Laplacian (add-one) smoothing

Given a corpus C, what is the probability of a word "New"?

$$P(New) = \frac{count(New)+1}{\sum_j count(w_j)+1}$$

Probability of a bigram

Given a corpus C, what is the probability of the phrase “New York”?

$$P(New) = \frac{count(New)}{\sum_j count(w_j)} \qquad P(York) = \frac{count(York)}{\sum_j count(w_j)}$$

We can’t just combine these probabilities

$$P(New, York)$$

We also care about the order of the words

$P(New)$ and the probability of $P(York\,|New)$

Probability of a bigram

Given a corpus C, what is the probability of the phrase "New York"?

$P(New)$ and the probability of $P(York\ |New)$

$$\text{P(New York)} = P(New)P(York\ |New)$$

$$P(New) = \frac{count(New)}{\sum_j count(w_j)}$$

$$P(York\ |\ New\ ) = \frac{count(New\ York)}{\sum_j count(New\ w_j)}$$

$$= \frac{count(New\ York)}{count(New)}$$

Probability of a bigram

Given a corpus C, what is the probability of the phrase “New York”?

$P(New)$ and the probability of $P(York\ |New)$

$$\text{P(New York)} = \frac{count(New)}{\sum_j count(w_j)} * \frac{count(New\ York)}{count(New)}$$

Language Modeling

Probability of a sentence based on bigrams

$$P(w_1 \ldots w_n) = \prod_i^n P(x_i | x_{i-1})$$

Probability of a sentence based on trigram

$$P(w_1 \ldots w_n) = \prod_i^n P(x_i | x_{i-1}, x_{i-2})$$

Collocation

Point-wise Mutual Information

$$PMI(x, y) = log \frac{P(x, y)}{P(x)P(y)}$$

$$PMI(w_1, w_2) = log \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

$$P(w_1, w_2) = P(w_2 \mid w_1)\, P(w_1)$$

$$PMI(w_1, w_2) = log \frac{P(w_2 \mid w_1)\, P(w_1)}{P(w_1)P(w_2)}$$

$$PMI(w_1, w_2) = log \frac{P(w_2 \mid w_1)\, \cancel{P(w_1)}}{\cancel{P(w_1)}P(w_2)}$$

Point-wise Mutual Information

$$PMI(x,y) = log\frac{P(y|x)}{P(y)}$$

$$PMI(w_1,w_2) = log\frac{P(w_2\,|w_1)}{P(w_2)}$$

How likely are we to see w_1 followed by w_2 normalized by how likely are we to see w_2 in general