LING 520: Computational Analysis of English

Semester: FALL '16

Instructor: Sowmya Vajjala

Iowa State University, USA

22 September 2016

Class outline

- morphological analysis - comments
- Probability overview
- ngrams and language models
- practice exercise

Stemming and Lemmatization - Comments

What are your observations about those different stemmers and lemmatizers in NLTK?

- How similar or different do those stemmers look in terms of the output you get?
- How is the lemmatizer different from stemmers?

Probability Overview

What is probability?

- Probability of an event is a mathematical way of predicting the chance that the event will occur in the given background.
- If I toss a coin once, what is the probability that I see a head?

What is probability?

- Probability of an event is a mathematical way of predicting the chance that the event will occur in the given background.
- If I toss a coin once, what is the probability that I see a head?
- Sample space (S): All possible occurrences of a event. In this coin toss example, there are two possibilities: head, tail.
- Event: each one of those possible occurrences can be one event. Having heads is one event, having tails is one event.

What is probability? - Some Formulae

- Probability is a number between 0 to 1.

What is probability? - Some Formulae

- Probability is a number between 0 to 1.
- If there are two possible events (E1,E2) in a sample space, P(E1) $+$ P(E2) $=$ 1.

What is probability? - Some Formulae

- Probability is a number between 0 to 1.
- If there are two possible events (E1,E2) in a sample space, $P(E1) + P(E2) = 1$.
- $P(S\text{-}E1) = 1\text{-}P(E1)$

What is probability? - Some Formulae

- Probability is a number between 0 to 1.
- If there are two possible events (E1,E2) in a sample space, $P(E1) + P(E2) = 1$.
- $P(S\text{-}E1) = 1\text{-}P(E1)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

What is probability? - Some Formulae

- Probability is a number between 0 to 1.
- If there are two possible events (E1,E2) in a sample space, P(E1) $+$ P(E2) $=$ 1.
- P(S-E1) $=$ 1-P(E1)
- P(A$\cup$B) $=$ P(A) $+$ P(B) - P(A$\cap$B)

What is probability? - Examples

- In our class, there are 12 or 13 people.
- If I have to randomly (and unbiasedly) pick one person to present on stemming now, what is the probability that it is Lena?

What is probability? - Examples

- In our class, there are 12 or 13 people.
- If I have to randomly (and unbiasedly) pick one person to present on stemming now, what is the probability that it is Lena? Answer: $1/13$
- What is the probability that it will be Lena or Kim?

What is probability? - Examples

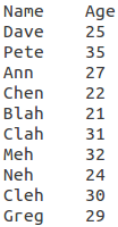
- In our class, there are 12 or 13 people.
- If I have to randomly (and unbiasedly) pick one person to present on stemming now, what is the probability that it is Lena? Answer: $1/13$
- What is the probability that it will be Lena or Kim? Answer: $1/13 + 1/13 = 2/13$ (it is a either or case). P(A∩B) is zero here.
- What is the probability that I pick either a second year ALT student or an American student?

What is probability? - Examples

- In our class, there are 12 or 13 people.
- If I have to randomly (and unbiasedly) pick one person to present on stemming now, what is the probability that it is Lena? Answer: 1/13
- What is the probability that it will be Lena or Kim? Answer: $1/13 + 1/13 = 2/13$ (it is a either or case). P(A$\cap$B) is zero here.
- What is the probability that I pick either a second year ALT student or an American student?
  Formula: P(2nd year ALT student) + P(American student) - P(2nd year ALT students who are americans).

What is probability? - Examples

- Look at this age distribution for 10 students:

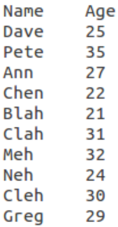


Name	Age
Dave	25
Pete	35
Ann	27
Chen	22
Blah	21
Clah	31
Meh	32
Neh	24
Cleh	30
Greg	29

- If I randomly pick one person, what is the probability that this person is below 30 years of age?

What is probability? - Examples

- Look at this age distribution for 10 students:

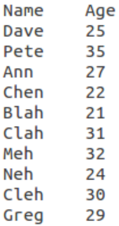


Name	Age
Dave	25
Pete	35
Ann	27
Chen	22
Blah	21
Clah	31
Meh	32
Neh	24
Cleh	30
Greg	29

- If I randomly pick one person, what is the probability that this person is below 30 years of age? Ans: 6/10
- If I randomly pick one person, what is the probability that it is Dave? what is the probability that this is not Dave?

What is probability? - Examples

- Look at this age distribution for 10 students:



Name	Age
Dave	25
Pete	35
Ann	27
Chen	22
Blah	21
Clah	31
Meh	32
Neh	24
Cleh	30
Greg	29

- If I randomly pick one person, what is the probability that this person is below 30 years of age? Ans: 6/10
- If I randomly pick one person, what is the probability that it is Dave? what is the probability that this is not Dave? Ans: 1/10 and 9/10.

Conditional probability

- Conditional probability is the probability of one event happening, when we know some other event has happened before.
- If one of my events is seeing 2 when I roll a die (E1), the other event is seeing an even number (E2), then, $P(E1|E2)$ is: $1/3$. Why??

Conditional probability

- Conditional probability is the probability of one event happening, when we know some other event has happened before.
- If one of my events is seeing 2 when I roll a die (E1), the other event is seeing an even number (E2), then, P(E1|E2) is: 1/3. Why??
- What is P(E1)?

Conditional probability

- Conditional probability is the probability of one event happening, when we know some other event has happened before.
- If one of my events is seeing 2 when I roll a die (E1), the other event is seeing an even number (E2), then, P(E1|E2) is: 1/3. Why??
- What is P(E1)? 1/6
- What is P(E2)?

Conditional probability

- Conditional probability is the probability of one event happening, when we know some other event has happened before.
- If one of my events is seeing 2 when I roll a die (E1), the other event is seeing an even number (E2), then, P(E1|E2) is: 1/3. Why??
- What is P(E1)? 1/6
- What is P(E2)? 3/6 i.e., 1/2
- What is P(E1, given E2)

Conditional probability

- Conditional probability is the probability of one event happening, when we know some other event has happened before.
- If one of my events is seeing 2 when I roll a die (E1), the other event is seeing an even number (E2), then, P(E1|E2) is: 1/3. Why??
- What is P(E1)? 1/6
- What is P(E2)? 3/6 i.e., 1/2
- What is P(E1, given E2) E2 has three possibilities (2,4,6). So, probability of getting 2 is 1/3.

Joint probability

- Probability that two events occur together. Represented as P(A,B) and is the same as P(B,A)
- So what is the difference between joint and conditional probability?

Joint probability

- Probability that two events occur together. Represented as P(A,B) and is the same as P(B,A)
- So what is the difference between joint and conditional probability?
- Useful example: https://goo.gl/9MVM78

Joint probability

- Probability that two events occur together. Represented as P(A,B) and is the same as P(B,A)
- So what is the difference between joint and conditional probability?
- Useful example: https://goo.gl/9MVM78
- Summary: Conditional probability is not commutative P(A|B) != P(B|A)
- P(A,B) = P(A|B)*P(B) = P(B|A)*P(A)

Bayes Theorem

- P(A | B) = P(B | A)* P(A) / P(B)
- Example: P(Tag|Word) = P(Word|Tag)*P(Tag)/P(Word)
- P(Word) and P(Tag) are probabilities of seeing the Word or Tag in your language corpus, independent of each other.
- What are P(Tag|Word) and P(Word|Tag)?

Bayes Theorem

- P(A | B) = P(B | A)* P(A) / P(B)
- Example: P(Tag|Word) = P(Word|Tag)*P(Tag)/P(Word)
- P(Word) and P(Tag) are probabilities of seeing the Word or Tag in your language corpus, independent of each other.
- What are P(Tag|Word) and P(Word|Tag)?
- P(Tag|Word) is the posterior probability. P(Word|Tag) is called the likelihood. P(Tag) is prior. P(Word) is evidence.

Ref: An explanation of this terminology: `https://goo.gl/bFwcHG`

Probability: Summary

- Probability is a number between 0 and 1.
- Joint probabilities are different from Conditional probabilities.
- P(A|B) != P(B|A)
- Bayes theorem: P(A | B) = P(B | A)* P(A) / P(B)

Note: I am only telling you what I want you to know to understand the rest of this class.

ngrams and language models

What are ngrams?

- Ngram is a token sequence of N words/tokens.
- Bigrams are two token sequences, trigrams are 3 token sequences etc.
- Ngram model: is a probabilistic model of language, that predicts the next word based on the previous words.
- Order is important. So, a sequence like "I am liking this class because ..." will get more probability in a language model for English than "this liking class I am because..."

Where are ngrams and ngram models useful?

1. In any task where we need to identify words in the presence of noisy input.
2. In speech recognition, to identify which of the similar sounding words is the word being spoken in the given context.
3. In handwriting recognition and optical character recognition, to disambiguate words or characters that are not clearly identified.
4. Machine translation: choosing from a possible set of word orderings between source and target translation.
5. Spelling and grammar correction: doing context sensitive spelling and grammar check.
6. POS tagging, natural language generation, predictive text input.. and many more.
7. Non modeling use: Doing corpus analysis.

Where may they fail?

- When there is a relatively free word-order language
- When there is a lot of word inflection
- When we need some deeper linguistic analysis (at the level of syntax, discourse etc)
- When there are tons of possible spelling variations possible.
- When you create your language model for one language, but use it for another language, or with totally new vocabulary etc.

Constructing ngram models

- Start with counting ngram sequences in a large corpora.
- goal: predict the next word based on some context. E.g., what is most likely to occur after ”Ngram is a sequence of ”
- In terms of probabilities:
  1. If ”ngram is a sequence of” is the context or history
  2. Let us say there is an english word ”whatever”
  3. ngram model should estimate P(”whatever”|”ngram is a sequence of”)
  4. How can we get this probability?

Probability of Language

- If we have a representative corpus of the style of language you want to see,
- If you have a program to count the frequency of occurrence ngrams of varying sizes,
- then, P(whatever|ngram is a sequence of) = C(ngram is a sequence of whatever| ngram is a sequence of)

Sounds pretty straight forward, doesn't it?

Probability of Language

- With this approach, if you have a large enough corpus, you can reliably estimate the probability of any sequence of words in a language.
- But, there are a few issues:
    1. Language is creative. New forms of word combinations keep coming up. No corpus can cover everything.

Probability of Language

- With this approach, if you have a large enough corpus, you can reliably estimate the probability of any sequence of words in a language.
- But, there are a few issues:
    1. Language is creative. New forms of word combinations keep coming up. No corpus can cover everything.
    2. Most of the sentences will have a zero probability, if there is one new word in it, or one unusual usage.

Probability of Language

- With this approach, if you have a large enough corpus, you can reliably estimate the probability of any sequence of words in a language.
- But, there are a few issues:
    1. Language is creative. New forms of word combinations keep coming up. No corpus can cover everything.
    2. Most of the sentences will have a zero probability, if there is one new word in it, or one unusual usage.
    3. Imagine writing a program for counting 2–10 ngrams and storing all this info, for such a large corpus!

Probability of Language

- With this approach, if you have a large enough corpus, you can reliably estimate the probability of any sequence of words in a language.
- But, there are a few issues:
  1. Language is creative. New forms of word combinations keep coming up. No corpus can cover everything.
  2. Most of the sentences will have a zero probability, if there is one new word in it, or one unusual usage.
  3. Imagine writing a program for counting 2–10 ngrams and storing all this info, for such a large corpus!
  4. .. and then imagine how to quickly retrieve this count information from those millions of ngrams you will have!

The chain rule of probability

- Here is one way of combining joint and conditional probabilities to estimate a sentence probability.
- If we assume a sentence S to be a sequence of words (w1, w2, ... wn), then, the probability of this sentence will be:
- P(S) = p(w1|beginning of sentence)*p(w2|w1)*p(w3|w1,w2)*p(w4|w1,w2,w3)*... ... ... *p(wn|w1,w2,w3...wn-1)
- This is perhaps a better representation, but is not making our job of counting ngrams any easier!
- One solution: markov assumption

The Markov assumption

- What is it?: It says instead of computing probabilities using entire history, we can use only the last few previous words.
- P(whatever|ngram is a sequence of) can be approximated as P(whatever|sequence of) if we choose a trigram model (2 words before current word).
- P(whatever) - unigram model. P(whatever|of) -bigram model .. and so on.

A Bigram Language Model

- This is the formula for getting bigram probabilities for one word: $\mathrm{P}(\mathrm{w}_n|\mathrm{w}_{n-1}) = \mathrm{C}(\mathrm{w}_{n-1},\mathrm{w}_n)/\Sigma_w(\mathrm{w}_{n-1}\mathrm{w})$

A Bigram Language Model

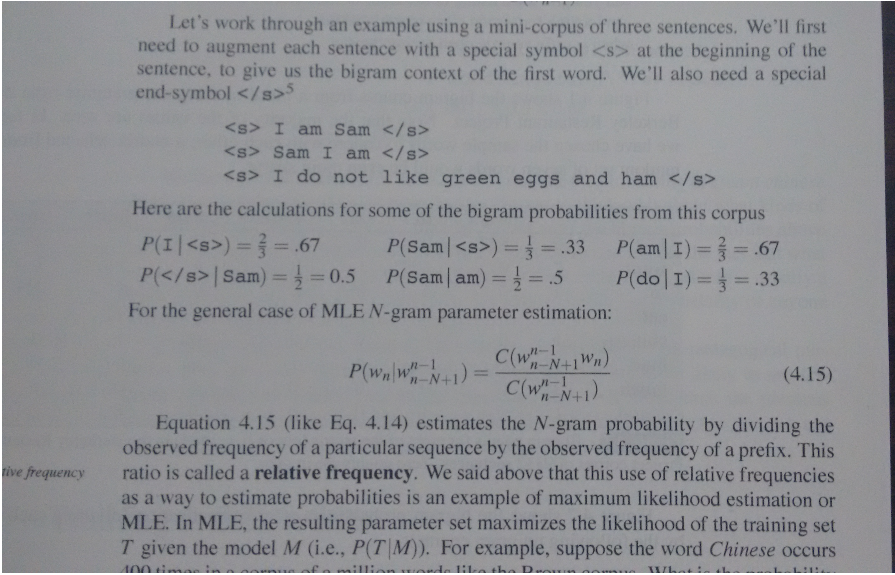
- This is the formula for getting bigram probabilities for one word: $\mathrm{P}(\mathrm{w}_n|\mathrm{w}_{n-1}) = \mathrm{C}(\mathrm{w}_{n-1},\mathrm{w}_n)/\Sigma_w(\mathrm{w}_{n-1}\mathrm{w})$
- However, $\Sigma_w(\mathrm{w}_{n-1}\mathrm{w})$ just means $\mathrm{C}(\mathrm{w}_{n-1})$. Why?

A Bigram Language Model

- This is the formula for getting bigram probabilities for one word: $P(w_n|w_{n-1}) = C(w_{n-1},w_n)/\Sigma_w(w_{n-1}w)$
- However, $\Sigma_w(w_{n-1}w)$ just means $C(w_{n-1})$. Why?
- The problem of several bigram combos being zero is still not resolved. Since our formula has a multiplication, this can potentially give zero probability to any given sentence even if there is only one bigram with zero probability. We will discuss how to address that part on tuesday.

An Example

Source: Section 4.2, Page 89 in J&M



Let's work through an example using a mini-corpus of three sentences. We'll first need to augment each sentence with a special symbol `<s>` at the beginning of the sentence, to give us the bigram context of the first word. We'll also need a special end-symbol `</s>`[5]

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>
```

Here are the calculations for some of the bigram probabilities from this corpus

$$P(\texttt{I}\,|\,\texttt{<s>}) = \tfrac{2}{3} = .67 \qquad P(\texttt{Sam}\,|\,\texttt{<s>}) = \tfrac{1}{3} = .33 \qquad P(\texttt{am}\,|\,\texttt{I}) = \tfrac{2}{3} = .67$$

$$P(\texttt{</s>}\,|\,\texttt{Sam}) = \tfrac{1}{2} = 0.5 \qquad P(\texttt{Sam}\,|\,\texttt{am}) = \tfrac{1}{2} = .5 \qquad P(\texttt{do}\,|\,\texttt{I}) = \tfrac{1}{3} = .33$$

For the general case of MLE N-gram parameter estimation:

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_n)}{C(w_{n-N+1}^{n-1})} \tag{4.15}$$

tive frequency

Equation 4.15 (like Eq. 4.14) estimates the N-gram probability by dividing the observed frequency of a particular sequence by the observed frequency of a prefix. This ratio is called a **relative frequency**. We said above that this use of relative frequencies as a way to estimate probabilities is an example of maximum likelihood estimation or MLE. In MLE, the resulting parameter set maximizes the likelihood of the training set T given the model M (i.e., $P(T|M)$). For example, suppose the word *Chinese* occurs 400 times in a corpus of a million words like the Brown corpus. What is the probability

Practice Exercise

Go to the following url, follow the python code snippets given there, and understand what is happening at each step. You can work in groups of 3. `https://goo.gl/U1ghTw`. I would expect one or two groups to give a summary of what they understood, in Tuesday's class.

Next Class

- Topics: continuation of language models
- ToDo:
  1. Read this article on Language Modeling in Python by Nitin Madnani (http://desilinguist.org/pdf/langmodel.pdf). We will start the next class with a discussion on this. Please do read.
  2. Watch Week 6 Lectures by Radev.