Introduction to Computation for the Humanities and Social Sciences

CS 3

Chris Tanner

Lecture 17

"No news is good news."
But *all* the news is even better*

Unless, it's current news, then it's all bad.

Lecture 17

- Language Models (Unigrams, Bigrams)
- APIs

Language Models

- a **Language Model** is a representation of the language used by a given entity (e.g., a person or a genre or any other well-defined author of text)

- Typically, these models are represented by statistics that are largely focused on just word counts

Text Analysis

```
[my, cat's, breath, smells, like, cat, ____]
```

Text generation

Language models

It all starts with a language model. A language model is at the core of many NLP tasks, and is simply a probability distribution over a sequence of words:

$$p(w_1, \ldots, w_n)$$

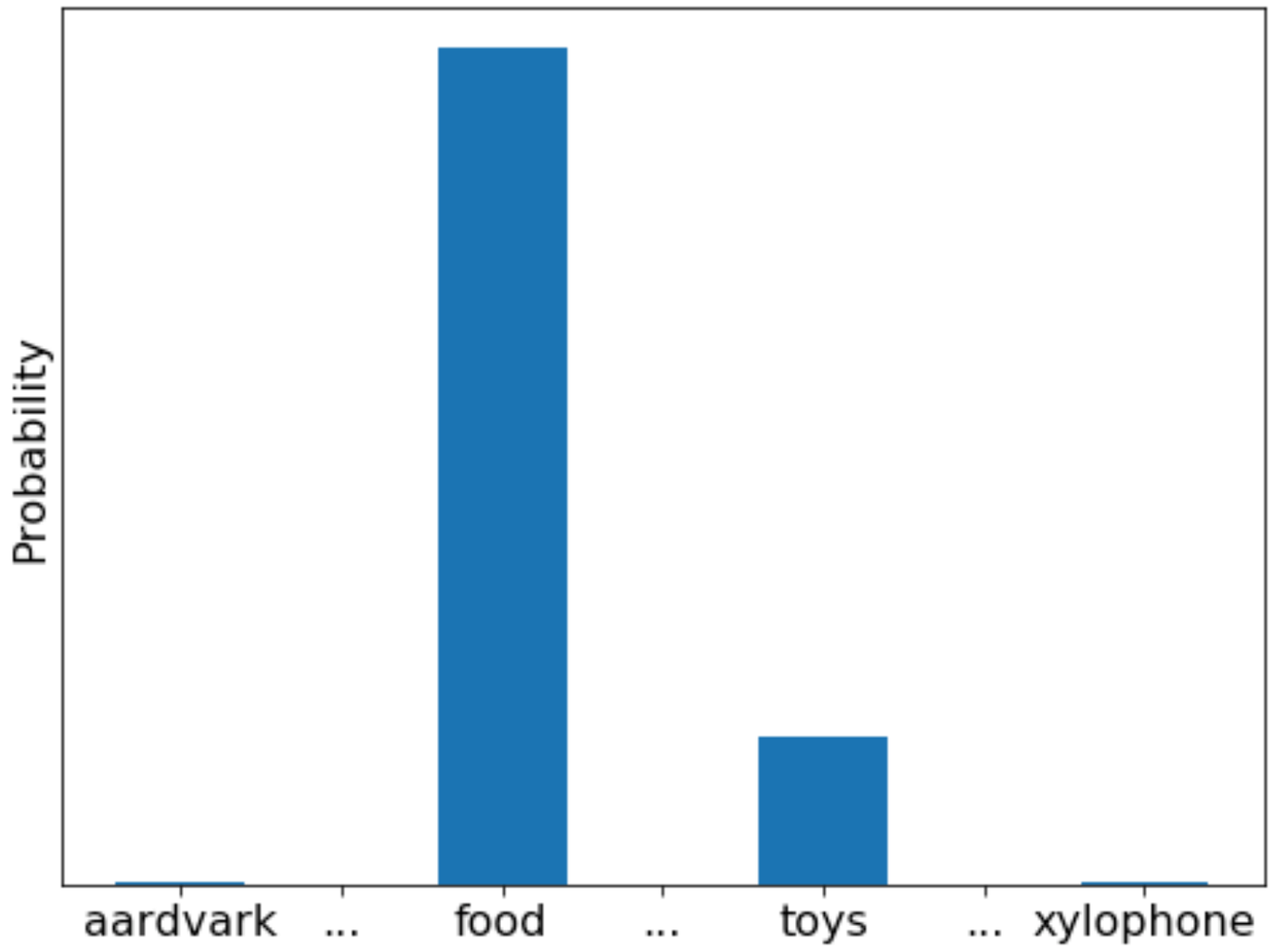
It can also be used to estimate the conditional probability of the next word in a sequence:

$$p(w_n \mid w_1, \ldots, w_{n-1})$$

https://medium.com/phrasee/neural-text-generation-generating-text-using-conditional-language-models-a37b69c7cd4b

Text Analysis

Let's assume we have the sequence `[my, cat's, breath, smells, like, cat, ____]` and we want to guess the final word. A language model would estimate the probability for every word in the vocabulary:



https://medium.com/phrasee/neural-text-generation-generating-text-using-conditional-language-models-a37b69c7cd4b

Keaton Patti
@KeatonPatti

Follow

I forced a bot to watch over 1,000 hours of lawyer commercials and then asked it to write a lawyer commercial of its own. Here is the first page.

R COMMERCIAL

FIRM LAW ROOM

YER stands next to a shelf with books. The boo
wide. They have eaten too many words.

LAWYER
Have you been hurt in an accidental car? Has the government sold your lungs without asking nicely? Are you Mesothelioma? Answer me!

awyer opens a briefcase. It's full of lemons,
ce fruit only lawyers may touch.

LAWYER (CONT'D)
If so, you can act entitled for money. I'll help. I graduated from lawn school and all my teachers were bitten by dogs.

scroll across bottom of the screen. These are
r takes: **UNFAIR STABBING, ILLEGAL SHOES, MUSIC IAN, SUE THE RAIN, DIVORCE YOUR TOILET, FAKE S**

LAWYER (CONT'D)
I have been a lawyer for over 35 weekends and I'm currently dating the Bill of Rights for fun.

ee the Bill of Rights. It's in love. The lawyer
k its heart. There's nothing we can do.

LAWYER (CONT'D)
Let me use it to send your asbestos to court. I will wear two suits and I promise to steal the judge's gavel for you.

lawyer opens up the jacket of his first suit. Mi
ls pour out. His promise has worth.

LAWYER (CONT'D)
My clients never go to jail town.

ee his past clients: a tornado, a tornado, a tor

LAWYER (CONT'D)
Remember, you don't pay any money unless you pay us money. Call for a free use of phone.

phone digits appear. It's your social security r

7:43 AM - 11 Oct 2018

LAWYER COMMERCIAL

INT. FIRM LAW ROOM

A LAWYER stands next to a shelf with books. The books are very wide. They have eaten too many words.

LAWYER
Have you been hurt in an accidental car? Has the government sold your lungs without asking nicely? Are you Mesothelioma? Answer me!

The lawyer opens a briefcase. It's full of lemons, the justice fruit only lawyers may touch.

LAWYER (CONT'D)
If so, you can act entitled for money. I'll help. I graduated from lawn school and all my teachers were bitten by dogs.

Words scroll across bottom of the screen. These are cases the lawyer takes: **UNFAIR STABBING, ILLEGAL SHOES, MUSIC TOO CANADIAN, SUE THE RAIN, DIVORCE YOUR TOILET, FAKE SONS.**

Language Models

How do we build such a model?

Unigram Language Models

- The most simple approach involves counting all of the individual words that an author uses (i.e., unigram counts).

- You've already done this before! **word_counts = {}**

- Then, you can construct probabilities by just dividing by the total # of words the author uses. e.g., maybe she/he uses the words:
  'reductionist' = 0.01% of the time
  'imperialism' = 0.04% of the time
  'capital' = 0.06% of the time
  etc

Unigram Language Models

- With these probabilities, you can do several fun things:
    - compare the model of one author to a different author
    - measure the likelihood that a given body of text was generated from a given user

Unigram Language Models

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.02%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.06%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.1%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.09%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.08%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.31%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

1.02%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

"C
mi
bo
ec
political discourse had grown too vitriolic"

If we assume each word is independent from another, we could just multiply every word's probability to come up w/ how likely the entire sequence came from the particular author

Unigram Language Models

Weaknesses with this unigram model?

Bi-gram Language Models

- The word order doesn't matter
- There is no context — doesn't look at any of the preceding words
- Instead, we could look at every pair of two consecutive words! That'll help some!

Bi-gram Language Models

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models — Let's first count all bigrams!

1

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models — Let's first count all bigrams!

1

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models — Let's first count all bigrams!

1

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models — Let's first count all bigrams!

1

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models — Let's first count all bigrams!

1

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models

For your Project #2, we only require you to build a dictionary of the bi-gram counts. You do not need to further turn those into probabilities.

Lecture 17

- Language Models (Unigrams, Bigrams)
- APIs

APIs

What are they?

- API stands for Application Programming Interface

- It's truly an interface — a mechanism — to use others' applications.

- Similar to using external libraries/packages, but instead of just using their provided collection of functions, we get to use their full-fledge app (aka program), which often outputs a lot of data for us to use.

What are they?

- Many large companies provide APIs so that others can use their application and data in their own programs.

    - Facebook statuses
    - Twitter tweets
    - Google Search Results
    - News stories!

News API

- We will use the News API, which allows us to access news stories from many different news outlets

- Go to https://newsapi.org/ and make an account

- `pip install newsapi-python`

- `from newsapi import NewsApiClient`

News API

- api = NewsApiClient(api_key=“your key goes here")
- top_headlines = api.get_top_headlines(q=‘trump')

```
{
    status: "ok",
    totalResults: 20,
  - articles: [
      - {
          - source: {
                id: "cnn",
                name: "CNN"
            },
            author: "Sophie Tatum, CNN",
            title: "CMS administrator says this year's scariest Halloween costume is
            'Medicare for all'",
            description: "The administrator of the nation's federal health insurance
            programs tweeted a Halloween-themed message bashing \"Medicare-for-all,\" a
            policy pushed by some Democrats.",
            url: https://www.cnn.com/2018/10/31/politics/seema-verma-halloween-tweet-
            medicare-for-all/index.html,
            urlToImage: https://cdn.cnn.com/cnnnext/dam/assets/161202121112-seema-
            verma-1122-super-tease.jpg,
            publishedAt: "2018-10-31T22:45:09Z",
            content: "Washington (CNN) The administrator of the nation's federal health
            insurance programs tweeted a Halloween-themed message bashing \"Medicare-
            for-all,\" a policy pushed by some Democrats. Seema Verma, who runs the
            Centers for Medicare and Medicaid Services, wrote … [+2320 chars]"
        },
      - {
          - source: {
                id: "the-hill",
                name: "The Hill"
            },
```

dictionary w/ 3 keys (“status”, “totalResults”, “articles”)


```
{
    status: "ok",
    totalResults: 20,
  - articles: [
      - {
          - source: {
                id: "cnn",
                name: "CNN"
            },
            author: "Sophie Tatum, CNN",
            title: "CMS administrator says this year's scariest Halloween costume is
            'Medicare for all'",
            description: "The administrator of the nation's federal health insurance
            programs tweeted a Halloween-themed message bashing \"Medicare-for-all,\" a
            policy pushed by some Democrats.",
            url: https://www.cnn.com/2018/10/31/politics/seema-verma-halloween-tweet-
            medicare-for-all/index.html,
            urlToImage: https://cdn.cnn.com/cnnnext/dam/assets/161202121112-seema-
            verma-1122-super-tease.jpg,
            publishedAt: "2018-10-31T22:45:09Z",
            content: "Washington (CNN) The administrator of the nation's federal health
            insurance programs tweeted a Halloween-themed message bashing \"Medicare-
            for-all,\" a policy pushed by some Democrats. Seema Verma, who runs the
            Centers for Medicare and Medicaid Services, wrote … [+2320 chars]"
        },
      - {
          - source: {
                id: "the-hill",
                name: "The Hill"
            },
```

dictionary w/ 3 keys (“status”, “totalResults”, “articles”)

“articles” -> list of dictionaries (each one pertains to one article)


```
{
    status: "ok",
    totalResults: 20,
  - articles: [
      - {
          - source: {
                id: "cnn",
                name: "CNN"
            },
            author: "Sophie Tatum, CNN",
            title: "CMS administrator says this year's scariest Halloween costume is
            'Medicare for all'",
            description: "The administrator of the nation's federal health insurance
            programs tweeted a Halloween-themed message bashing \"Medicare-for-all,\" a
            policy pushed by some Democrats.",
            url: https://www.cnn.com/2018/10/31/politics/seema-verma-halloween-tweet-
            medicare-for-all/index.html,
            urlToImage: https://cdn.cnn.com/cnnnext/dam/assets/161202121112-seema-
            verma-1122-super-tease.jpg,
            publishedAt: "2018-10-31T22:45:09Z",
            content: "Washington (CNN) The administrator of the nation's federal health
            insurance programs tweeted a Halloween-themed message bashing \"Medicare-
            for-all,\" a policy pushed by some Democrats. Seema Verma, who runs the
            Centers for Medicare and Medicaid Services, wrote … [+2320 chars]"
        },
      - {
          - source: {
                id: "the-hill",
                name: "The Hill"
            },
```

```
{
    status: "ok",
    totalResults: 20,
  - articles: [
      - {
          - source: {
                id: "cnn",
                name: "CNN"
            },
            author: "Sophie Tatum, CNN",
            title: "CMS administrator says this year's scariest Halloween costume is
            'Medicare for all'",
            description: "The administrator of the nation's federal health insurance
            programs tweeted a Halloween-themed message bashing \"Medicare-for-all,\" a
            policy pushed by some Democrats.",
            url: https://www.cnn.com/2018/10/31/politics/seema-verma-halloween-tweet-
            medicare-for-all/index.html,
            urlToImage: https://cdn.cnn.com/cnnnext/dam/assets/161202121112-seema-
            verma-1122-super-tease.jpg,
            publishedAt: "2018-10-31T22:45:09Z",
            content: "Washington (CNN) The administrator of the nation's federal health
            insurance programs tweeted a Halloween-themed message bashing \"Medicare-
            for-all,\" a policy pushed by some Democrats. Seema Verma, who runs the
            Centers for Medicare and Medicaid Services, wrote … [+2320 chars]"
        },
      - {
          - source: {
                id: "the-hill",
                name: "The Hill"
            },
```


dictionary w/ 3 keys (“status”, “totalResults”, “articles”)

“articles” -> list of dictionaries (each one pertains to one article)

each article dictionary has many keys, including source, title, content, etc

LAB TIME