Language & Technology

Lecture 4: Word-Based Models

Thomas Graf

Stony Brook University
lin120@thomasgraf.net

Counting Words

- Here's something nobody would ever call fun: **counting words in a text**.
- But word counts are easy for computers, and they are surprisingly useful:
  - culturomics
  - *adsense* (online ad placement)
  - word meaning
  - authorship attribution
- So let's see how that works.

Assembling a Corpus

Corpus a collection of texts

Example

- Collected works of Shakespeare
- All tweets between 2012 and 2016
- Google books
- Wall Street Journal (WSJ) corpus
- Brown corpus

- All applications of word counting models need **lots of text**.
- Corpora are collections of texts.
- Often they have been prepared for easy use with computers.

Splitting a Corpus into Word Lists

- Corpora usually consist of plain text files.
  think txt rather than doc or pdf
- We can read in a text file and use a regular expression to break the string into a list of words.
  This is called **tokenization**.

```
1  string = "The sun shone, having no alternative, on the nothing new."
2  tokens = re.findall("\w+", str.lower(string))
3  print(tokens)
4  >>> ["the", "sun", "shone", "having", "no", "alternative", "on",
5       "the", "nothing", "new"]
```

- But what do we do with that list?

Counting Words

- It is easy to count how often each word occurs.

```
1  from collections import Counter
2  counts = Counter(tokens)
3  print(counts)
4  >>> {"the": 2, "alternative": 1, "having": 1, "new": 1, "no": 1,
5       "nothing": 1, "on": 1, "shone": 1, "sun": 1}
```

- More technically: for each word type we count its word tokens.

word type a word of a given language

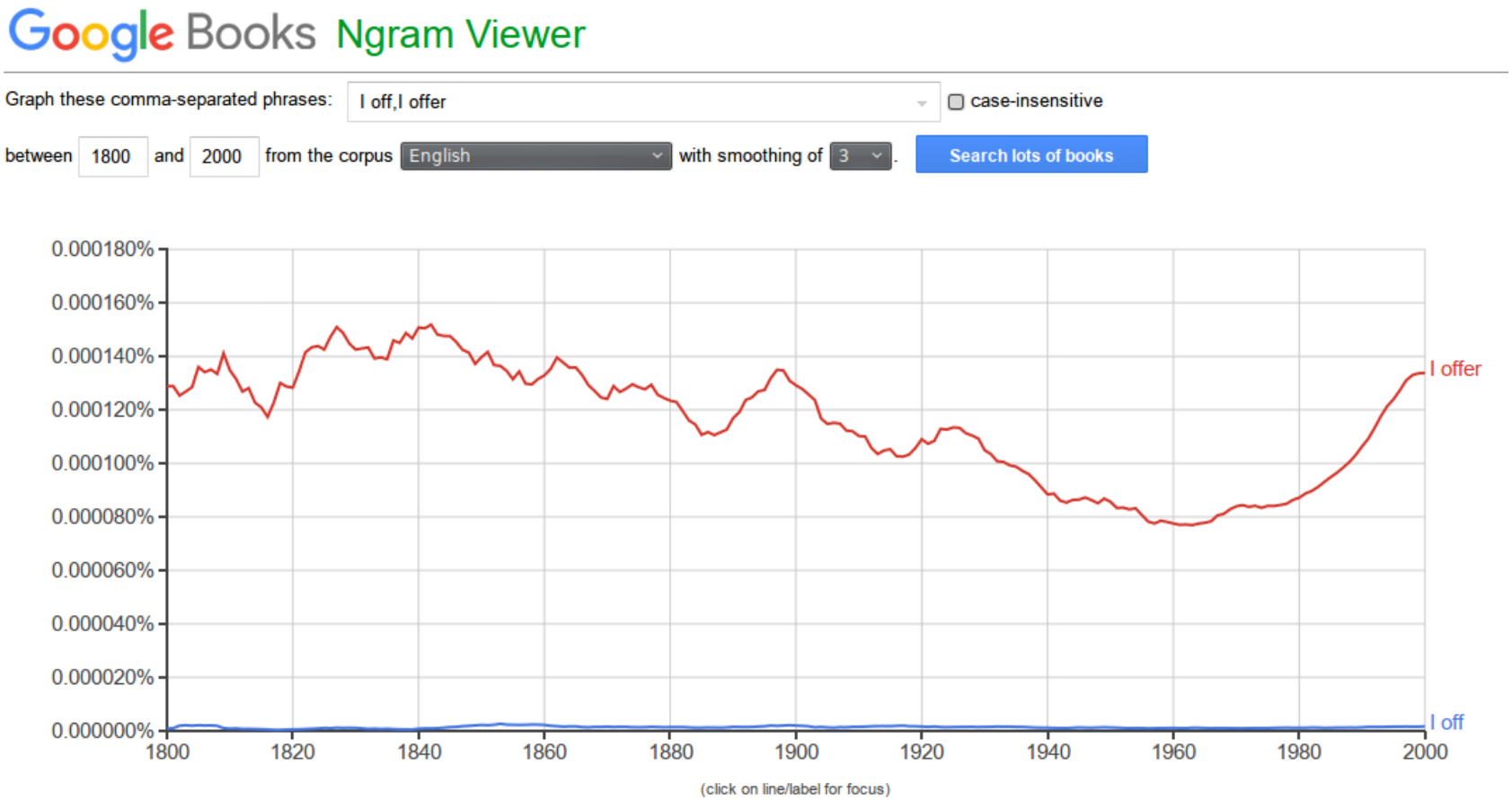
word token instance of the word in a text

Some Terminology: Unigrams and n-Grams

- When we only count words in isolation,
  we are building a **unigram** model.
- Unigram models are a special case of **n-gram** models,
  which count sequences of words.
- We will learn more about n-gram models at a later point.

A Word Count Application: Culturomics

- **Culturomics** is the quantitative study of cultural trends with the help of corpora.
- How does the frequency of specific words change over time?
- Let's try it ourselves with the Google n-gram viewer!



Careful: Don't Misinterpret the Data

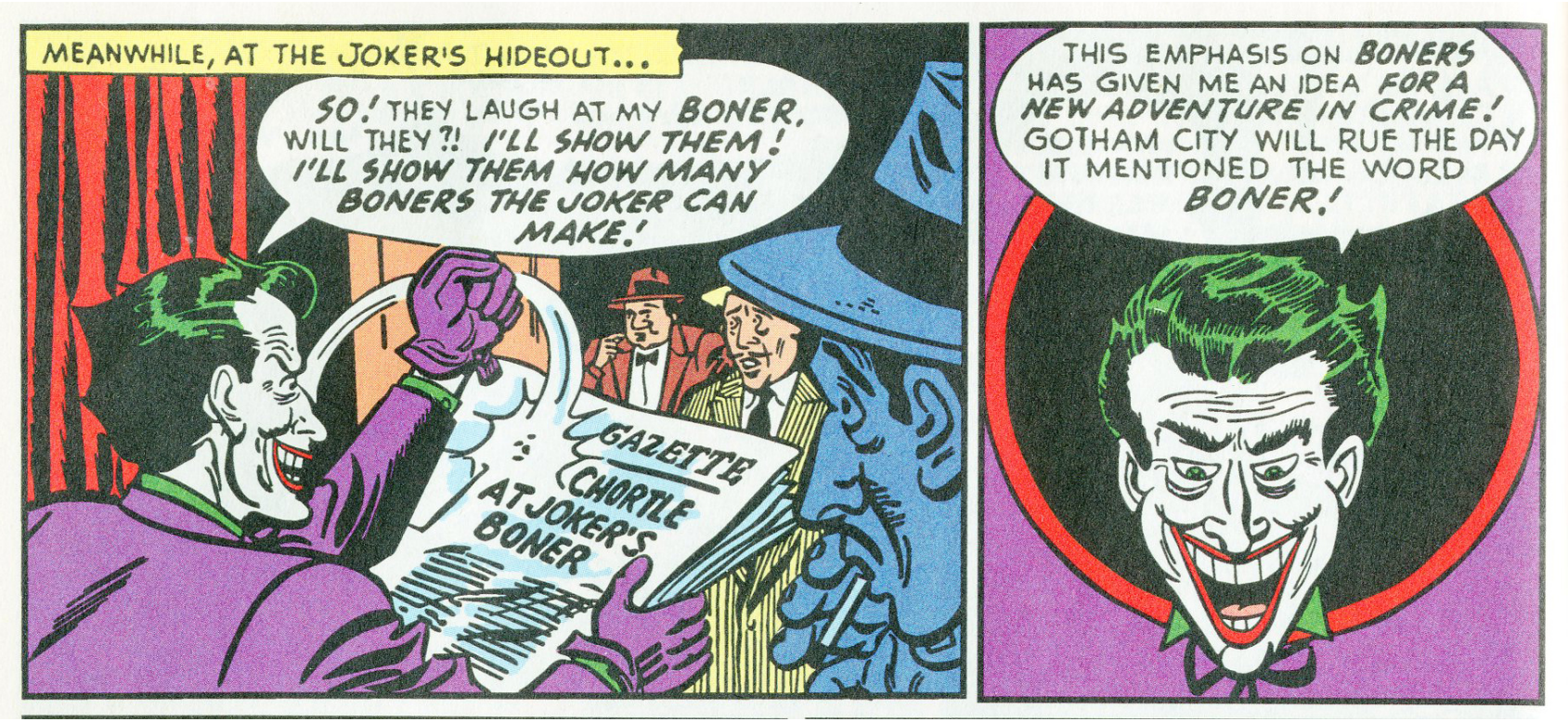
- The connection between language and culture is very indirect.
- Words have many meanings, we often don't know which meaning is tracked by frequencies.

Example: Tracking Racism

- **black** could also be color
- **chink** could be used for "crack", "gap"
- **redskin** could be used for a fan of the sports team

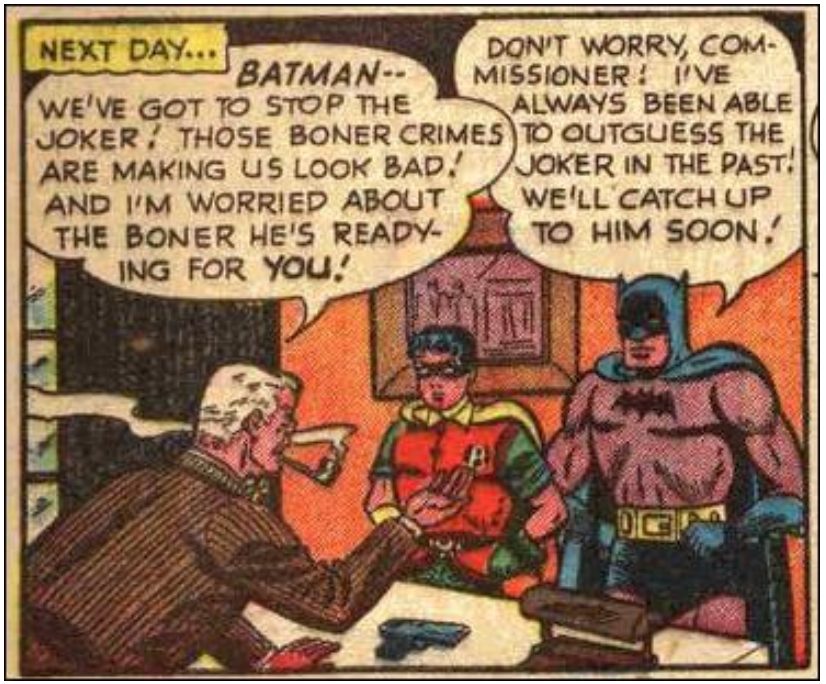
- Additional problem: word meanings change over time

A Real-World Example of Language Change

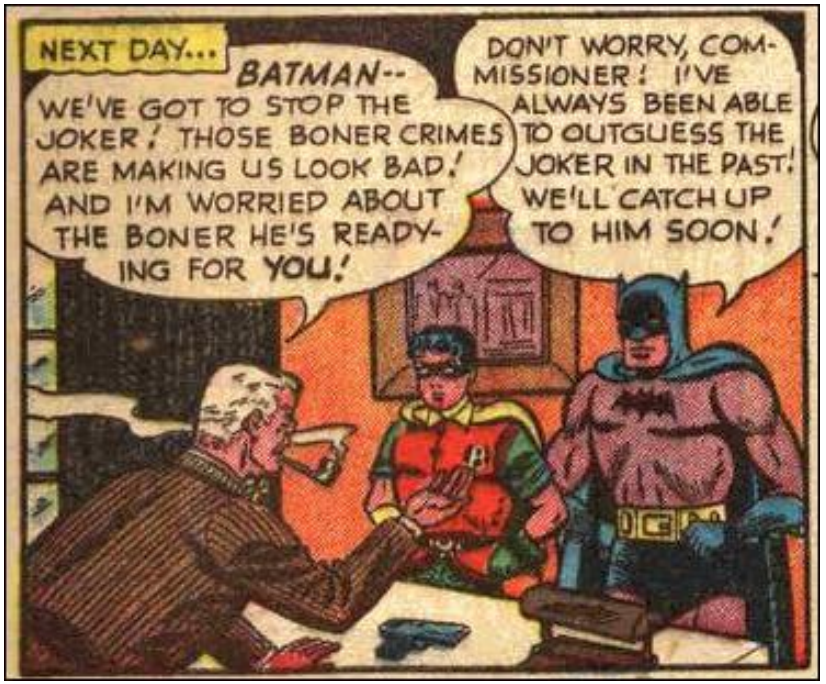


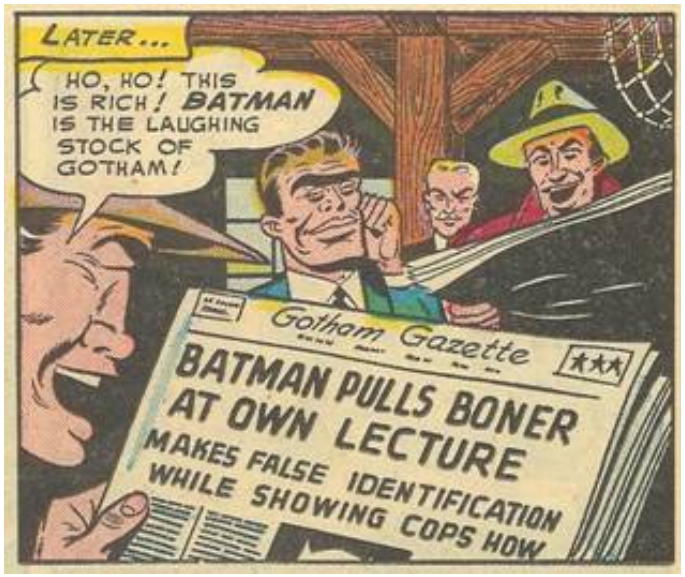
And There's More Where That Came From

And There's More Where That Came From



And There's More Where That Came From





A Remark on Google Books

- Google does tons of stuff with unigram and n-gram models.
- The more data they have, the better the models.
- When Google Books launched, people thought it was about
  - extending Google's search business to books,
  - creating a digital library,
  - digital preservation.
- Few realized it was all about getting
  more data for their models!

Some Success Stories of Culturomics

Culturomics is still a young field, but there are interesting results:

- **Fukushima shift**
  Analysis of 5 million news paper articles reveals public shift on nuclear power after Fukushima accident.
- **Twitter forecast**
  Public sentiment can be reconstructed and traced via Twitter.
- **Lexical point of no return**
  If a word does not disappear 30 to 50 years after its introduction, it is very unlikely to disappear at a later point.

Reasons to be Sceptical

- As every new tool, culturomics is likely to be used incorrectly.
- In particular, claims about public opinion are dubious.
- **Methodological issue:**
  How could we reliably falsify such claims?
- **Linguistic issue:** The occurrence of certain words in a sentence tells us very little about its meaning.

(1) a. **Twitter post on sunny day:** What a lovely day!
    b. **Twitter post on rainy day:** What a lovely day!

(2) a. I'm asking you and Bill not to vote for Trump.
    b. I'm asking you and not Bill to vote for Trump.
    c. I'm not asking you and Bill to vote for Trump.

Culturomics: The Jury is Still Out

- Culturomics is a great tool as long as the questions are simple and restricted:
  - frequency of words and phrases over time
  - trending topics
- But unigram/n-gram models are too limited for meaning.
- They make reliable claims about people's opinions/attitudes only if their shortcomings can be offset by enough data.
- No proof so far that large corpora solve the linguistic issues

The Sceptic's Guide to Culturomics

- Whether people talk about topic X is easy to study.
- What people think about topic X is much more difficult.

Culturomics: The Jury is Still Out

- Culturomics is a great tool as long as the questions are simple and restricted:
  - frequency of words and phrases over time
  - trending topics
- But unigram/n-gram models are too limited for meaning.
- They make reliable claims about people's opinions/attitudes only if their shortcomings can be offset by enough data.
- No proof so far that large corpora solve the linguistic issues

The Sceptic's Guide to Culturomics

- Whether people talk about topic X is easy to study.
- What people think about topic X is much more difficult.

Learning More: Computational Sociology at Stony Brook

- Jason Jones
- SOC 330 Media and Society

Application 2: Predicting the Success of Novels

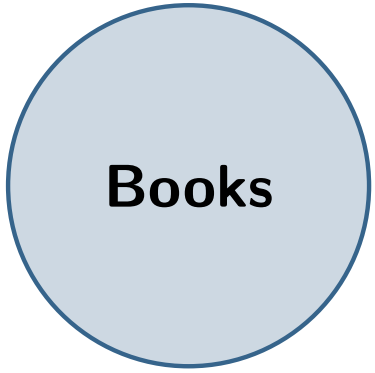
Success with Style: Using Writing Style to Predict the Success of Novels

- Study conducted here at Stony Brook by Vikas Ashok, Song Feng, and Yejin Choi
- **The Big Insight**
  Unigram models are surprisingly accurate at predicting the success of novels.
- **What it is NOT About**
  - computers writing successful novels
  - advice for becoming a successful author

Yejin Choi

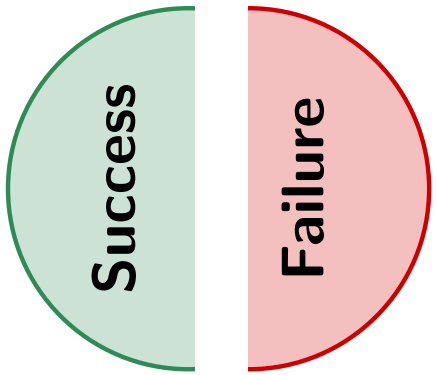
Methodology in Detail

1. collect books from Project Gutenberg
2. annotate them as (un)successful
   mixture of financial turnout and critical evaluation
3. split annotated books into two groups
   - **Training set**
     used to train the model/learn frequencies
   - **Test set**
     used to test quality of the trained model
4. determine unigram frequencies in training set and establish statistical correlation to success of books
5. get predictions of model for books in test set
6. compare predictions to actual success/failure



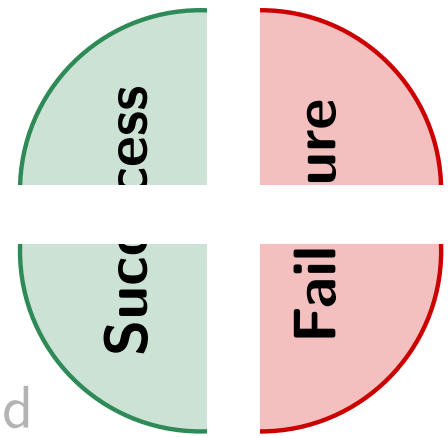
Methodology in Detail

1. collect books from Project Gutenberg
2. annotate them as (un)successful
   mixture of financial turnout and critical evaluation
3. split annotated books into two groups
   - **Training set**
     used to train the model/learn frequencies
   - **Test set**
     used to test quality of the trained model
4. determine unigram frequencies in training set and establish statistical correlation to success of books
5. get predictions of model for books in test set
6. compare predictions to actual success/failure



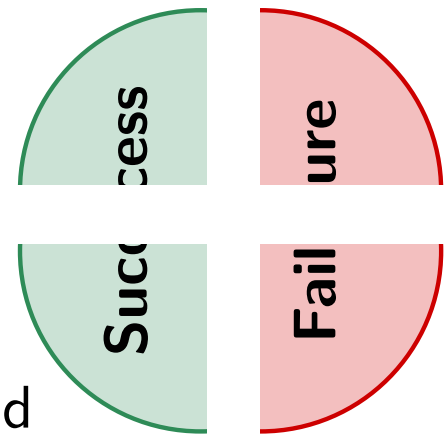
Methodology in Detail

1. collect books from Project Gutenberg
2. annotate them as (un)successful
   mixture of financial turnout and critical evaluation
3. split annotated books into two groups
   - **Training set**
     used to train the model/learn frequencies
   - **Test set**
     used to test quality of the trained model
4. determine unigram frequencies in training set and establish statistical correlation to success of books
5. get predictions of model for books in test set
6. compare predictions to actual success/failure



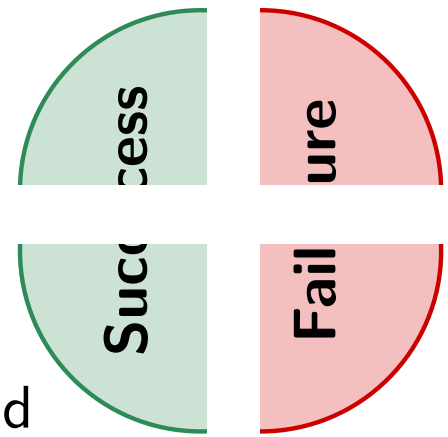
Methodology in Detail

1. collect books from Project Gutenberg
2. annotate them as (un)successful
   mixture of financial turnout and critical evaluation
3. split annotated books into two groups
   - **Training set**
     used to train the model/learn frequencies
   - **Test set**
     used to test quality of the trained model
4. determine unigram frequencies in training set and establish statistical correlation to success of books
5. get predictions of model for books in test set
6. compare predictions to actual success/failure



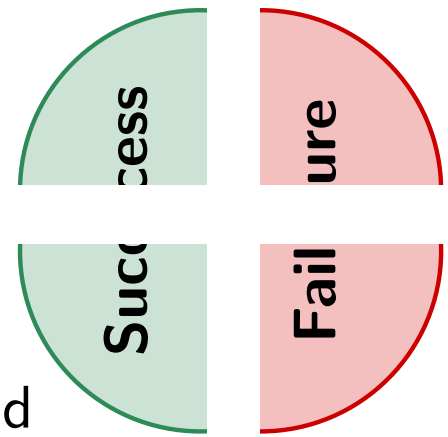
Methodology in Detail

1. collect books from Project Gutenberg
2. annotate them as (un)successful
   mixture of financial turnout and critical evaluation
3. split annotated books into two groups
   - **Training set**
     used to train the model/learn frequencies
   - **Test set**
     used to test quality of the trained model
4. determine unigram frequencies in training set and establish statistical correlation to success of books
5. get predictions of model for books in test set
6. compare predictions to actual success/failure



Methodology in Detail

1. collect books from Project Gutenberg
2. annotate them as (un)successful
   mixture of financial turnout and critical evaluation
3. split annotated books into two groups
   - **Training set**
     used to train the model/learn frequencies
   - **Test set**
     used to test quality of the trained model
4. determine unigram frequencies in training set and establish statistical correlation to success of books
5. get predictions of model for books in test set
6. compare predictions to actual success/failure



Findings

- Unigram model made right prediction for $\approx$**75%** of books
- More complicated models only marginally better ($\approx +2\%$)
- All models perform very badly on sci-fi and history.
- All models perform much better on romance and adventure.
- Successful books have a higher occurrence of *and* and thought-oriented verbs like *recognize* and *remember*.

The Big Question

- Those are interesting findings.
- But why do we find this? What does it mean?

Findings

- Unigram model made right prediction for $\approx$**75%** of books
- More complicated models only marginally better ($\approx +2\%$)
- All models perform very badly on sci-fi and history.
- All models perform much better on romance and adventure.
- Successful books have a higher occurrence of *and* and thought-oriented verbs like *recognize* and *remember*.

The Big Question

- Those are interesting findings.
- But why do we find this? What does it mean?

What it Doesn't Mean

- Stand-up comedian **Dave Gorman**'s show *Modern Life is Goodish* features what he calls found poems.
- **Idea:** collect outrageously stupid online comments into a poem
- In that spirit, here's some user comments from the *Telegraph*.

What it Doesn't Mean [cont.]

Don't you think if there was a way to analyse what made a best-seller, authors would have figured it out long before a bunch of scientists with nothing better to do? There is not, and never will be, a way to write a best-selling novel. This comes from an author of fifteen years and a dozen novels.

What it Doesn't Mean [cont.]

Don't you think if there was a way to analyse what made a best-seller, authors would have figured it out long before a bunch of scientists with nothing better to do? There is not, and never will be, a way to write a best-selling novel. This comes from an author of fifteen years and a dozen novels.

The Misunderstanding

- This is not about writing successful books.
- The study is about predicting success of written books.
- The model is too abstract to be turned into writing advise.
  you cannot write with unigram frequencies

What it Doesn't Mean [cont.]

This algorithm has an issue. they are analyzing books which are SOLD! Not books which are on stock and should be sold. I means: How can a reader know how many "adverbs" or "adjectives" there are into a book, if the book is NOT yet bought by him?? Thus once the book is ought, it's sold. And certainly the reader cannot bring the book back to store and claim the money back, because he doesn't like the book.

What it Doesn't Mean [cont.]

This algorithm has an issue. they are analyzing books which are SOLD! Not books which are on stock and should be sold. I means: How can a reader know how many "adverbs" or "adjectives" there are into a book, if the book is NOT yet bought by him?? Thus once the book is ought, it's sold. And certainly the reader cannot bring the book back to store and claim the money back, because he doesn't like the book.

The Misunderstanding

- *Nitpick 1*: you can return books
- *Nitpick 2*: reviews and word-to-mouth drive sales
- *Real issue*
  - sold books are not part of the model at all
  - this is a categorization task, categories happen to be success/failure
  - the notion of success is not purely based on what sold

What it Doesn't Mean [cont.]

*These nitwits wanted their algorithm to discover whether a book (presumably in manuscript form) would be a "commercial success" — so they used it to back-predict the "commercial success" of classics from Project Gutenberg? Good God! What made them think that the classics even *were* commercial successes [. . .]*

What it Doesn't Mean [cont.]

> *These nitwits wanted their algorithm to discover whether a book (presumably in manuscript form) would be a "commercial success" — so they used it to back-predict the "commercial success" of classics from Project Gutenberg? Good God! What made them think that the classics even *were* commercial successes [. . .]*

The Misunderstanding

- Again, this is only about predicting the right category.
- The authors call this category "success", and they define it in a reasonable manner (commercial and/or critical success).
- How well this lines up with what you consider success is irrelevant.
- "Back-prediction" is an essential part of model testing.

What it Doesn't Mean [cont.]

Perhaps the reason they found common threads in the successful novels is because they were well written....duh! Did we really need a computer algorithm to tell us that?

What it Doesn't Mean [cont.]

Perhaps the reason they found common threads in the successful novels is because they were well written....duh! Did we really need a computer algorithm to tell us that?

The Misunderstanding

- What does it mean to be well-written?
- How do you formalize it for a computer?
- The model has nothing to say about the quality of the writing style.

What it Doesn't Mean [cont.]

Such an algorithm would be impossible for a book written by one of the masters of English literature. It's only possible for junk books.

What it Doesn't Mean [cont.]

Such an algorithm would be impossible for a book written by one of the masters of English literature. It's only possible for junk books.

The Misunderstanding

- Many of the books on Project Gutenberg are literary classics.
- Why would this only work for junk books?
  The n-gram model can be used with any text.

What it Doesn't Mean [cont.]

same guy as before, in a follow-up post:

> *I know what a computer algorithm is and it might be possible to "write a book" by this means if the book's "formula" was very simple indeed. But writing a great imaginative work of literature is not a step by step mechanical procedure and series of calculations by which a desired end is accomplished. That should be obvious to the meanest intelligence.*

What it Doesn't Mean [cont.]

same guy as before, in a follow-up post:

> *I know what a computer algorithm is and it might be possible to "write a book" by this means if the book's "formula" was very simple indeed. But writing a great imaginative work of literature is not a step by step mechanical procedure and series of calculations by which a desired end is accomplished. That should be obvious to the meanest intelligence.*

The Misunderstanding

- This is not about writing successful books.
- In particular, it is not about computers writing books.

What it Doesn't Mean [cont.]

The western world is now full of people who don't have jobs, but have employment. There is hardly a week goes by when some "scientist" after years of research states the bleeding obvious. Meanwhile over in China they have the spanners and hammers out building things. They make lots of money, and we borrow it so that coco the clown can come up with pointless drivel like this.

What it Doesn't Mean [cont.]

> *The western world is now full of people who don't have jobs, but have employment. There is hardly a week goes by when some "scientist" after years of research states the bleeding obvious. Meanwhile over in China they have the spanners and hammers out building things. They make lots of money, and we borrow it so that coco the clown can come up with pointless drivel like this.*

The Misunderstanding

o tempora, o mores

So What Does it Mean?

- If we view the collection of successful books at a group level, there are statistical trends regarding unigram frequencies.
- These trends are pronounced enough that they are a good predictor for whether a specific book belongs to that group.
- But these trends do not necessarily explain why books are successful.
- **Correlation is not causation!**

Soap Box Slide: Public Science Writing

- In defense of the commenters, the Telegraph article does a bad job summarizing the paper.
- This is a common theme with public science writing: they overhype and distort in their quest for clicks.

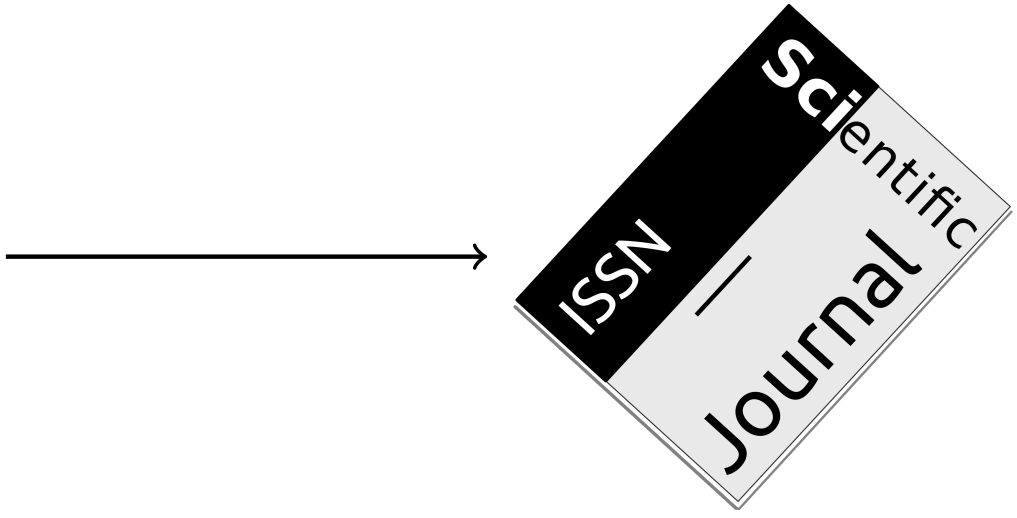
Moral of the Story

- Don't believe scientific news in mainstream media.
- Seek out specialized outlets:
  - science blogs
  - Quanta magazine
- Read the paper yourself (just the abstract and conclusion will already rectify the worst misconceptions).
- If the paper is behind a paywall
  - lobby congress to fund open access initiatives,
  - use `sci-hub.cc` (evil, evil piracy)

Soap Box Slide: The Telephone Game of Science Reporting

Soap Box Slide: The Telephone Game of Science Reporting

Soap Box Slide: The Telephone Game of Science Reporting



Soap Box Slide: The Telephone Game of Science Reporting

Soap Box Slide: The Telephone Game of Science Reporting

Soap Box Slide: Group-Level and Individual-Level

- group-level predictions $\neq$ individual-level predictions

Example

- Math tests: men, as a group, outperform women, as a group.
- But the correlation between group and individual is weak.
- Gender is a bad predictor of an individual's math skills.

- Many correlations only hold reliably at the group level (in particular those related to cognitive tasks).
- When they are strong enough to make for a useful predictor for individuals, that should grab your attention.