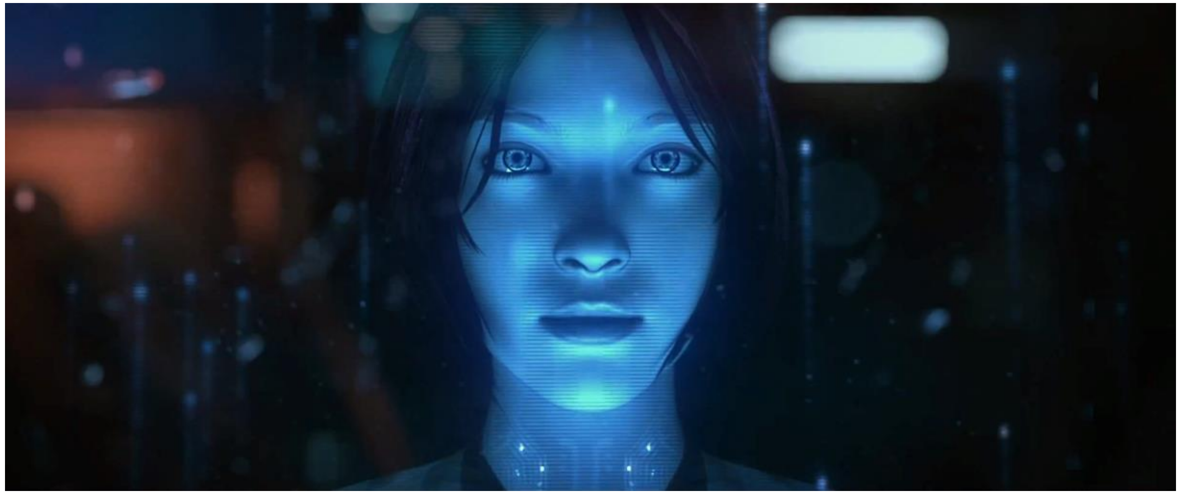
Basic Natural Language Processing

Why NLP?

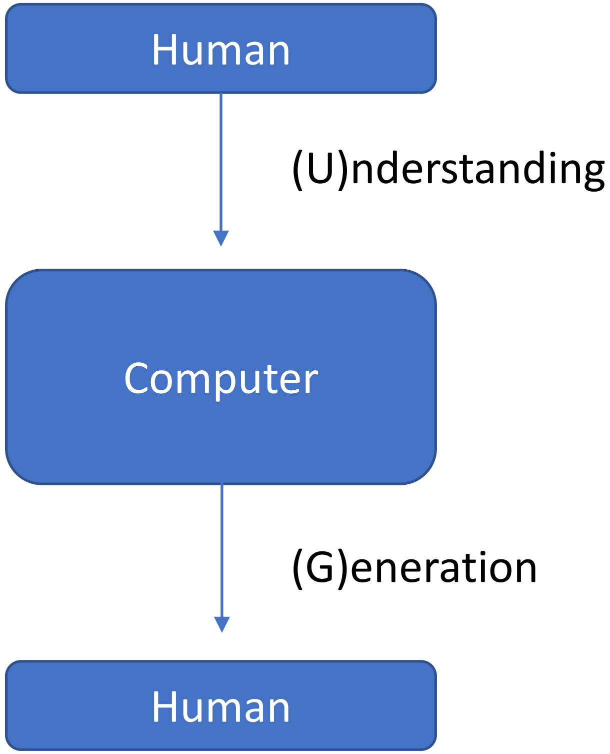
- Understanding Intent
  - Search Engines
- Question Answering
  - Azure QnA, Bots, Watson
- Digital Assistants
  - Cortana, Siri, Alexa
- Translation Systems
  - Azure Language Translation, Google Translate
- News Digest
  - Flipboard, Facebook, Twitter
- Other uses
  - Pollect, Crime mapping, Earthquake prediction



Understanding human language is hard

NLP requires inputs from :

- Linguistics
- Computer Science
- Mathematics
- Statistics
- Machine Learning
- Psychology
- Databases



THE KEY: Changing uncertainty to certainty

I	am	changing	this	sentence	to	numbers
1	2	3	4	5	6	7

“Vectorizing”

You	are	changing	too	many	sentences!
8	?	3	?	9	?

Remember: There is no ambiguity with numbers!

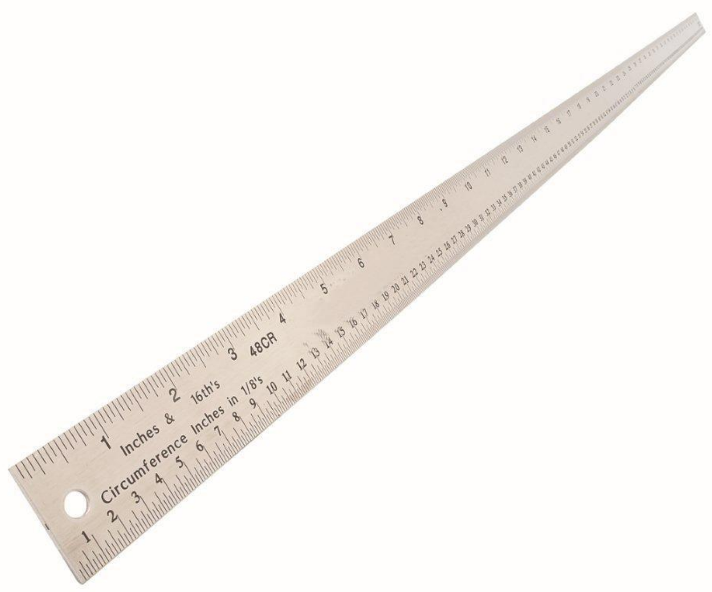
Challenges in NLP: Syntax vs. Semantics

- Syntax:
  - Lamb a Mary had little

- Semantics:
  - Merry hat hey lid tell lam
  - Colorless orange liquid
  - Address, number, resent

Challenges in NLP: Ambiguity pt 1

- CC Attachment
  - I like swimming in warm lakes and rivers
- Ellipsis and Parallelism
  - I gave the Steven a shovel and Joseph a ruler
- Metonymy
  - Sydney is essential to this class
- Phonetic
  - My toes are getting number
- Pp Attachment
  - You ate spaghetti with meatballs / pleasure / a fork / Jillian /



Challenges in NLP: Ambiguity pt 2

- Referential
  - Sharon complimented Lisl. She had been kind all day.
- Reflexive
  - Brandon brought himself an apple
- Sense
  - Julia took the math quiz
- Subjectivity
  - Karen believes that the Economy will stay strong
- Syntactic
  - Call a dentist for Wayne

Challenges in NLP: Others

- Parsing N-grams:
  - United States of America
  - Hot dog
- Typos
  - John Hopkins vs Johns Hopkins
- Non-standard language
  - (208)929-6136 vs 208-929-6136
  - Cause = because
- SARCASM
  - I *love* rotting apples

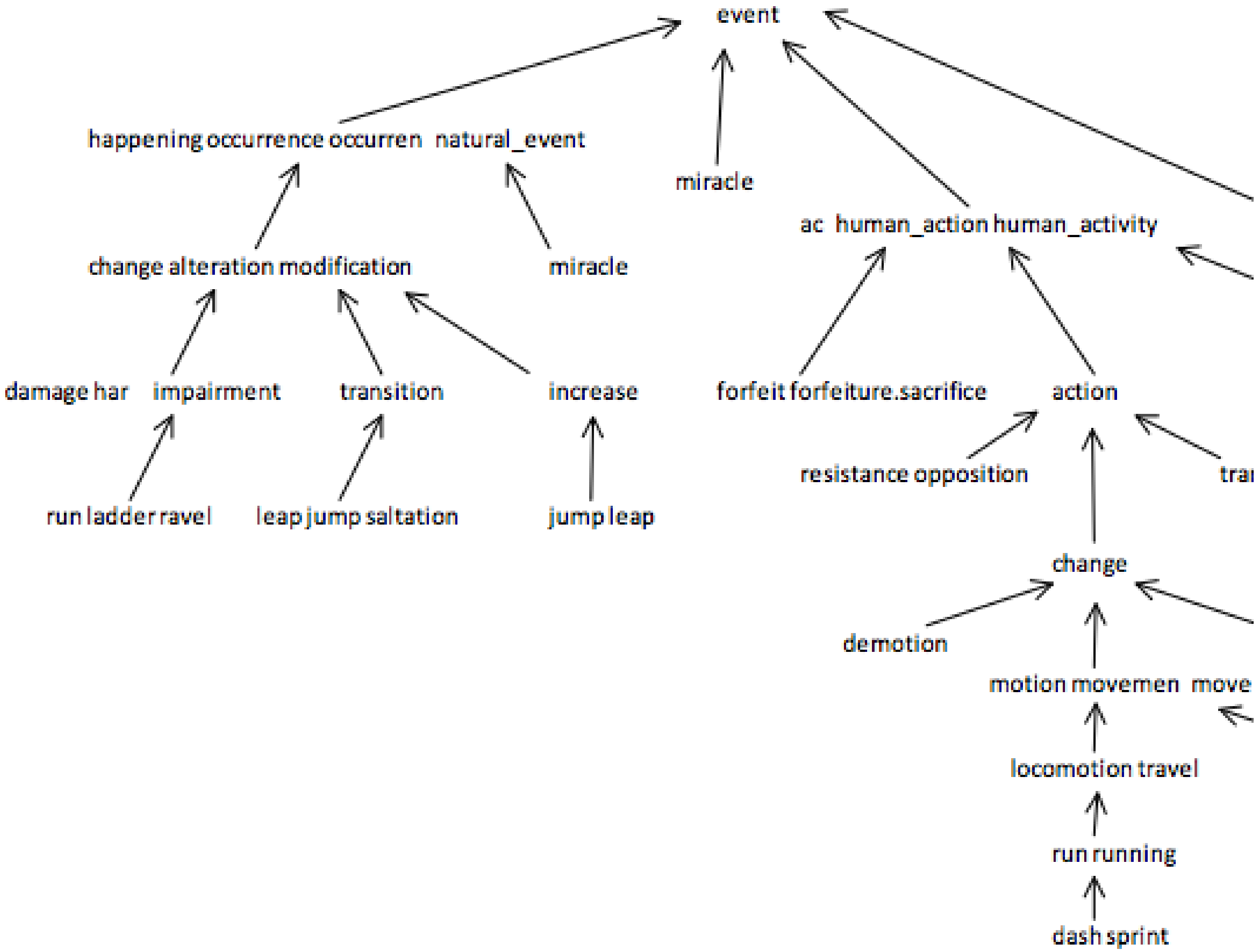
Edit Distance: How we Spellcheck

- Can reference box above, left, or diagonal up-left
- If letter matches, +0
- If letter doesn’t match, +1
- Score is the box at the bottom-right

		S	**T**	**R**	**E**	**N**	**G**	**T**	**H**
	0	**1**	2	3	4	5	6	7	8
T	1	1	**1**	2	3	4	5	5	6
R	2	2	2	**1**	2	3	4	5	6
E	3	3	3	2	**1**	2	3	4	5
N	4	4	4	3	2	**1**	**2**	**3**	4
D	5	5	5	4	3	2	2	3	**4**

Semantic Relationships

- Measuring how words are related to each other.
- Birdcage will be more similar to Dog Kennel than it will be to Bird
- Many different systems to draw out semantic relationships, but ‘Wordnet’ is one of the most commonly used
- Similarity metric:
- Sim(V,W) = - ln(pathlength(V,W))
- Sim(Run, Miracle) would be = -ln(7)



Preprocessing: Stopwords and punctuation

Why we want to get rid of them?

- “And”, “If”, “But”, “.”, “,”
- Will almost ALWAYS be your most significant words
- Tells you nothing about what’s going on

Don’t get rid of them if you are focused on Natural Language Generation!

Preprocessing: Porter’s Algorithm

Measure:

- A ‘**measure’** of a word is an indication of how many syllables are in it.
- Consonants = ‘C’, Vowels = ‘V’
- Every sequence of ‘VC’ is counted as +1
- Intellectual = (VC)C(VC)C(VC)CV(VC) = 4

Stemming:

- Strip a word down to its barest form
- Ex: ‘Alleviation’ – $\underbrace{\text{‘ation’ + ‘ate’}}_{\text{Transformational Rule}}$ = ‘Alleviate’

Stemming: Sample Rules

- If m>0:
  - Lies -> li
    - Abilities = Abiliti
  - Ational -> ate
    - National = National
    - Recreational = recreate
  - Sses -> ss
    - Sunglasses = sunglass
  - Biliti -> ble
    - Abiliti = able

Stemming: Example

- Original Word: “Computational”
  - Computational – ‘ational’ + ‘ate’ = Computate
  - Computate – ‘ate’ = Comput
- Final Word: “Comput”

- Original Word: “Computer”
  - Computer – ‘er’ = Comput
- Final Word: “Comput”

Sentence Boundary Recognition

Problems with things like Dr., A.M., U.S.A.

Use a decision tree to estimate the boundary

Features:

- Punctuation
- Formatting
- Fonts
- Spaces
- Capitalization
- Known Abbreviations

N-Gram Modeling

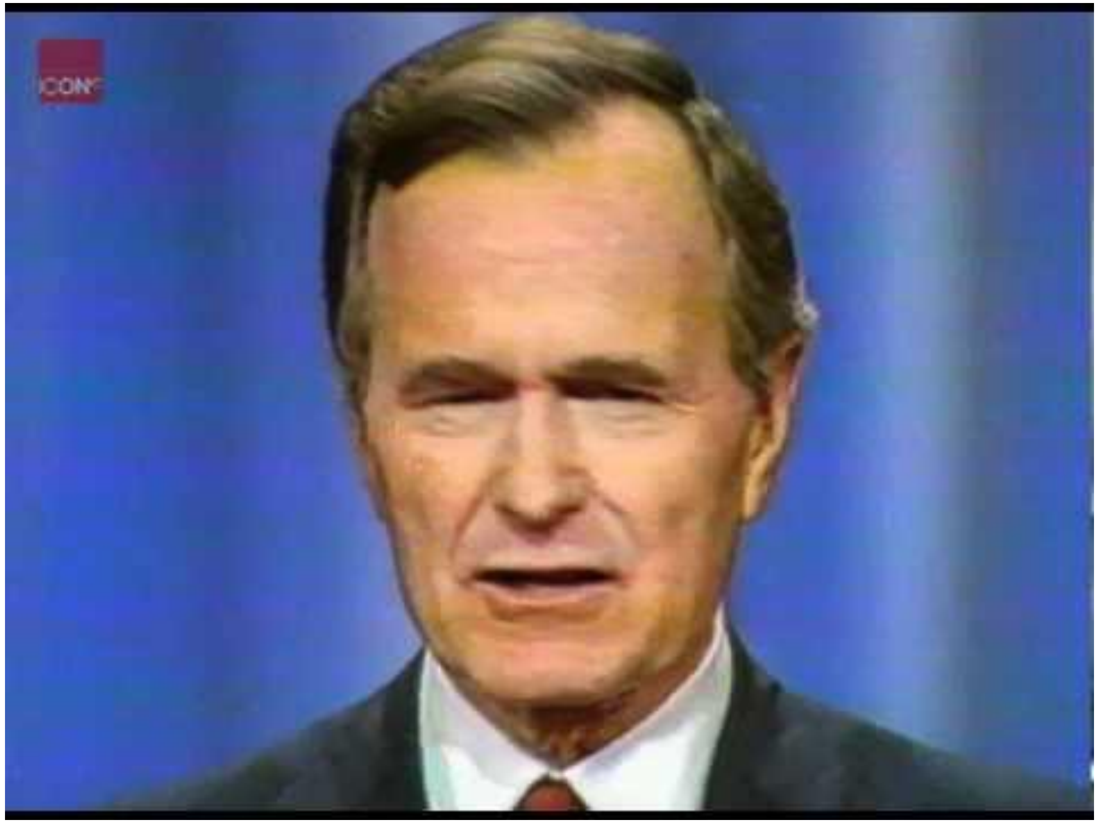
Words that have a separate meaning when combined with other words

The best way to highlight the importance of context

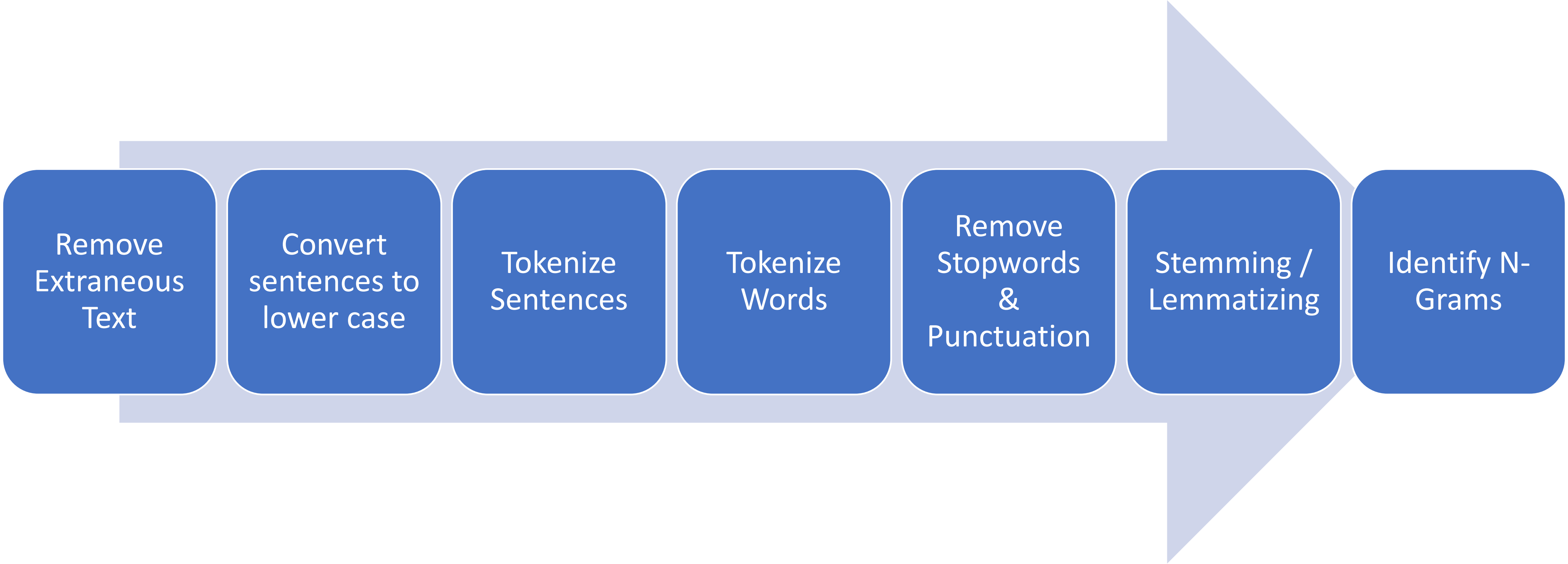
Examples:

- Unigram: Apple
- Bigram: Hot Dog
- Trigram: George Bush Sr.

I’ll meet you in Times {?????}



Preprocessing Checklist



Remove Extraneous Text

Convert sentences to lower case

Tokenize Sentences

Tokenize Words

Remove Stopwords & Punctuation

Stemming / Lemmatizing

Identify N-Grams

Words to Numbers

- Corpus creation
  - Create a library of all words in original dataset

- Vectorizing
  - Changing words to numbers
  - Often a raw count

- TFIDF
  - Term Frequency / Inverse Document Frequency
  - Example:
    - "This" mentioned 3 times in a given review, but the review has 27 words in it
    - Tfidf = 3 / 27 = 1/9

Bayes Theorem

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

Predicting the next { … }

Example from Charles Dickens:

- P(“Darnay looked at Dr. Manette”)
- Use maximum likelihood estimates for the n-gram probabilities
  - Unigram: P(w) = c(w)/V
  - Bigram: P(w1 | w2) = c(w1,w2)/c(w2)

- Values

  - P(“Darnay”) = 533 / 598633 = .00089
  - P(“looked”|”Darnay”) = 3 / 676 = .0044
  - P(“at|looked”) = 77 / 312 = .247
  - P(“Dr. Manette” | “at”) = 2 / 4512 = .000443

- Bigram probability

  - P(“Darnay looked at Dr. Manette”) = 4.28 * e^-10

- P(“at Dr. Manette Darnay looked”) = 0

The Bag of Words Approach

- P(Positive Review | Words Contained)
- Look at the unordered words of a document to determine underlying characteristics
- Coffee reviews with the word ‘bean’ tend to be far more positive
- Common in sentiment and feature analysis