|  |  |  |
| --- | --- | --- |
|  | Anthony Olivieri  Syracuse iSchool  Applied Data Science | |
|  | |  |

Portfolio

And Language Modeling

# Introduction

Background

Study of the similarities of documents by statistical analysis.

# Problem Specification

Goals

To gain a better understanding of language used in two disparate documents. The documents chosen are Lewis Carroll's *Alice's Adventures in Wonderland* and Walt Whitman's *Leaves of Grass*. The two were chosen because they were published in the same year, but exist in different categories: one a nonsensical fantasy in prose, the other a collection of poems from Civil War America.

The questions to be answered

* What are the differences between texts?
* Which words and bigrams were most common in each?

# Data

The gutenberg corpus files "whitman-leaves.txt," and "carroll-alice.txt" were brought in from the nltk library.

# Processing

Lowercase, Stopped, Tokenized

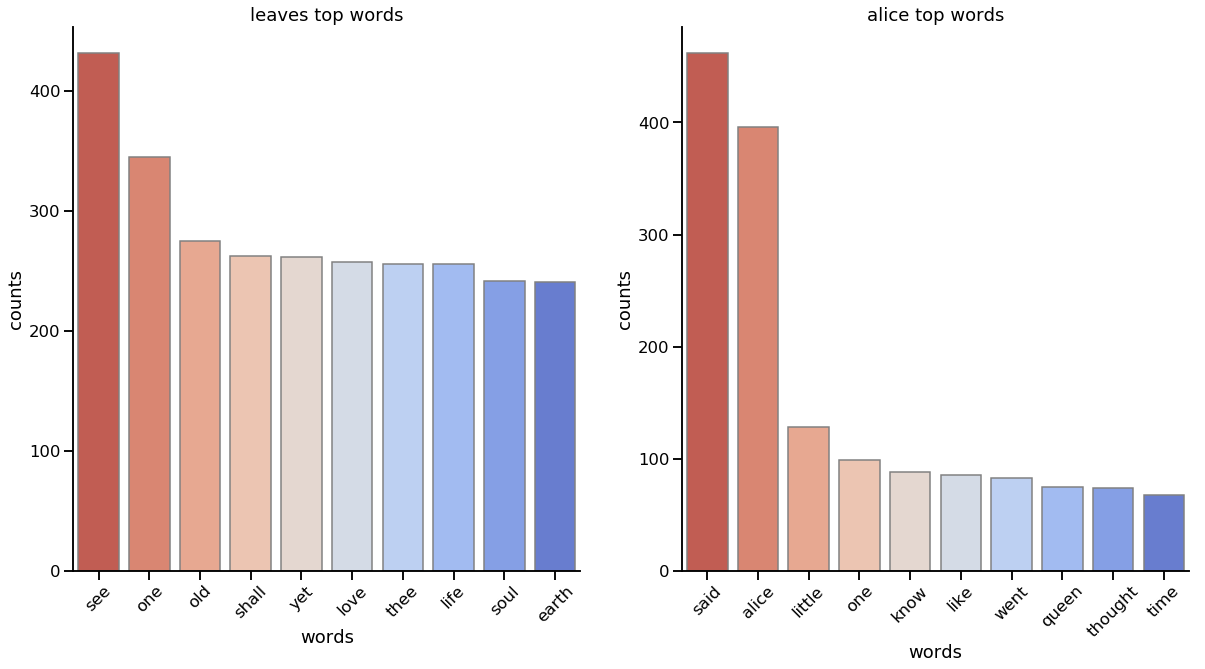
The texts were lowered in case, checked against stop-words, all punctuation was removed and they were tokenized. This created a list of lowercase words. The stopwords list was expanded to include "O" and "as" because running frequencies on the texts favored these two words. These are not necessarily descriptive of the genre or text so they were removed.

Frequency Distribution and Bigrams Scoring

Frequency distributions and bigrams were created. A PMI of 9 was used to reduce repetitions which were not indicative of the genre, but of the text itself. For example ('walt', 'whitman').

# Evaluation

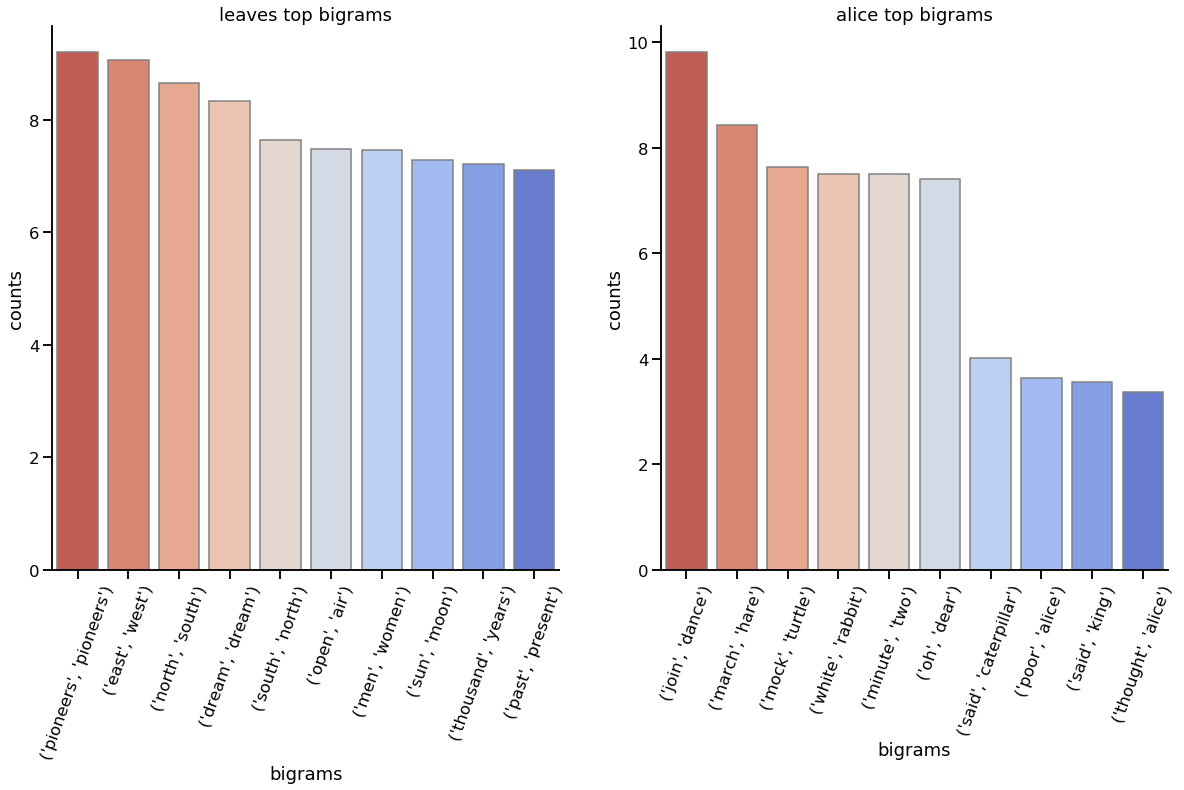
Top Words



"Alice" and "Queen" break the top 10 for Alice's Adventure in Wonderland, while no pronouns make it into Leaves of Grass. Alice's top words have more verbs "said, know, like, went, thought" than Leaves "see, shall, love". The scale of the adjectives in each is interesting, "little" makes it into the Alice top 10 and "earth" as well as "old" make it into Leaves.

Both drop off after the first two most popular words "see, one" for Leaves and "said Alice" for Alice, meaning the author uses the top spot very often. Alice, however, drops steeply after the first two while Leaves makes a more gradual downturn. Leaves also has much higher scores for places 3 through 10.

Top Bigrams



The higher scores for the bigrams that round out the top ten are here too, Alice drops off after ("oh", "dear"). Directions ("east", "west") and ("north", "south") as well as time, ("past", "present"), and ("thousand", "years") are represented in Leaves while Alice focuses on characters ("march", "hare"), ("white", "rabbit"), ("said", "caterpillar"), ("said", "alice"), ("said", "king"), and ("thought", "alice"). In whole Alice's bigrams describe characters and their actions while Leaves' bigrams describe a grandiose landscape in place and time.

This an intuitive result for a character-driven adventure and a tome of poetry in a wild country at war with itself. It may also speak to the size and development of each nation (England, and the US). England being a densely-populated and industrialized island and the US being a large open wilderness. The depth-of-focus for each author contained by the environment.

|  |  |
| --- | --- |
| Leaves Frequency | Alice Frequency |
| ('see', 432),  ('one', 345),  ('old', 275),  ('shall', 263),  ('yet', 262),  ('love', 258),  ('thee', 256),  ('life', 256),  ('soul', 242),  ('earth', 241),  ('man', 232),  ('thy', 226),  ('long', 224),  ('night', 212),  ('thou', 209),  ('day', 185),  ('men', 185),  ('every', 180),  ('death', 175),  ('know', 174),  ('come', 170),  ('time', 160),  ('great', 160),  ('upon', 156),  ('sea', 151),  ('ever', 149),  ('world', 149),  ('many', 141),  ('hear', 137),  ('face', 130),  ('like', 129),  ('never', 128),  ('hand', 126),  ('give', 125),  ('body', 124),  ('pass', 119),  ('look', 118),  ('good', 118),  ('young', 117),  ('last', 116),  ('well', 116),  ('go', 114),  ('land', 112),  ('rest', 109),  ('us', 109),  ('eyes', 107),  ('sing', 105),  ('air', 104),  ('nothing', 104),  ('women', 103) | ('said', 462),  ('alice', 396),  ('little', 128),  ('one', 99),  ('know', 88),  ('like', 85),  ('went', 83),  ('queen', 75),  ('thought', 74),  ('time', 68),  ('see', 67),  ('king', 62),  ('turtle', 58),  ('began', 58),  ('mock', 56),  ('hatter', 56),  ('quite', 55),  ('gryphon', 54),  ('think', 53),  ('way', 53),  ('say', 51),  ('much', 51),  ('head', 50),  ('first', 50),  ('thing', 49),  ('voice', 48),  ('go', 48),  ('rabbit', 47),  ('got', 45),  ('looked', 45),  ('never', 45),  ('get', 44),  ('mouse', 42),  ('duchess', 42),  ('round', 41),  ('tone', 40),  ('came', 40),  ('dormouse', 40),  ('great', 39),  ('back', 37),  ('well', 37),  ('two', 37),  ('cat', 36),  ('march', 34),  ('large', 33),  ('last', 33),  ('long', 32),  ('come', 31),  ('moment', 31),  ('put', 31) |

|  |  |
| --- | --- |
| Leaves Bigrams Raw | Alice Bigrams Raw |
| (('men', 'women'), 0.0008),  (('man', 'woman'), 0.0005),  (('day', 'night'), 0.0005),  (('young', 'men'), 0.0004),  (('pioneers', 'pioneers'), 0.0004),  (('every', 'one'), 0.0003),  (('old', 'age'), 0.0003),  (('young', 'man'), 0.0003),  (('long', 'long'), 0.0003),  (('thee', 'thy'), 0.0002),  (('north', 'south'), 0.0002),  (('old', 'man'), 0.0002),  (('open', 'air'), 0.0002),  (('time', 'space'), 0.0002),  (('life', 'death'), 0.0002),  (('thee', 'thou'), 0.0002),  (('east', 'west'), 0.0002),  (('every', 'day'), 0.0002),  (('every', 'thing'), 0.0002),  (('whole', 'earth'), 0.00018),  (('young', 'old'), 0.00018),  (('body', 'soul'), 0.00017),  (('look', 'upon'), 0.00017),  (('new', 'world'), 0.00017),  (('past', 'present'), 0.00017),  (('death', 'death'), 0.00015),  (('hand', 'hand'), 0.00015),  (('one', 'side'), 0.00015),  (('sun', 'moon'), 0.00015),  (('thousand', 'years'), 0.00015),  (('woman', 'man'), 0.00015),  (('dream', 'dream'), 0.00014),  (('far', 'far'), 0.00014),  (('let', 'others'), 0.00014),  (('let', 'us'), 0.00014),  (('little', 'child'), 0.00014),  (('one', 'another'), 0.00014),  (('south', 'north'), 0.00014),  (('thee', 'soul'), 0.00014),  (('western', 'sea'), 0.00014),  (('comes', 'back'), 0.00012),  (('face', 'face'), 0.000121),  (('male', 'female'), 0.000121),  (('old', 'men'), 0.000121),  (('old', 'young'), 0.000121),  (('one', 'love'), 0.000121),  (('one', 'one'), 0.000121),  (('soul', 'soul'), 0.000121),  (('thou', 'hast'), 0.000121),  (('walt', 'whitman'), 0.000121) | ((('said', 'alice'), 0.01),  (('mock', 'turtle'), 0.004),  (('march', 'hare'), 0.002),  (('said', 'king'), 0.002),  (('thought', 'alice'), 0.002),  (('alice', 'said'), 0.001),  (('said', 'hatter'), 0.001),  (('white', 'rabbit'), 0.001),  (('said', 'mock'), 0.001),  (('said', 'caterpillar'), 0.001),  (('said', 'gryphon'), 0.001),  (('said', 'duchess'), 0.001),  (('alice', 'thought'), 0.001),  (('said', 'cat'), 0.001),  (('said', 'queen'), 0.001),  (('minute', 'two'), 0.001),  (('king', 'said'), 0.0009),  (('poor', 'alice'), 0.0009),  (('said', 'march'), 0.0008),  (('alice', 'looked'), 0.0007),  (('alice', 'replied'), 0.0007),  (('alice', 'went'), 0.0007),  (('join', 'dance'), 0.0007),  (('oh', 'dear'), 0.0007),  (('said', 'dormouse'), 0.0007),  (('said', 'mouse'), 0.0007),  (('like', 'said'), 0.0006),  (('little', 'thing'), 0.0006),  (('poor', 'little'), 0.0006),  (('alice', 'began'), 0.0006),  (('alice', 'felt'), 0.0006),  (('alice', 'much'), 0.0006),  (('alice', 'quite'), 0.0006),  (('course', 'said'), 0.0006),  (('cried', 'alice'), 0.0006),  (('good', 'deal'), 0.0006),  (('indeed', 'said'), 0.0006),  (('know', 'said'), 0.0006),  (('thing', 'said'), 0.0006),  (('alice', 'like'), 0.0005),  (('beautiful', 'soup'), 0.0005),  (('beg', 'pardon'), 0.0005),  (('dance', 'join'), 0.0005),  (('golden', 'key'), 0.0005),  (('great', 'hurry'), 0.0005),  (('hare', 'said'), 0.0005),  (('little', 'door'), 0.0005),  (('queen', 'said'), 0.0005),  (('said', 'nothing'), 0.0005),  (('set', 'work'), 0.0005) |

|  |  |
| --- | --- |
| Leaves Bigrams PMI | Alice Bigrams PMI |
| (('1492', 'awaken'), 15.954969),  (('16th', '17th'), 15.954969),  (('19-20', '1881'), 15.954969),  (('43', 'despise'), 15.954969),  (('72d', '73d'), 15.954969),  (('abase', 'abased'), 15.954969),  (('abject', 'louse'), 15.954969),  (('abroad', 'swift-rising'), 15.954969),  (('abstinence', 'falsehood'), 15.954969),  (('acceptances', 'rejections'), 15.954969),  (('accidents', 'rebuffs'), 15.954969),  (('accouche', 'accouchez'), 15.954969),  (('accouchez', 'rot'), 15.954969),  (('accoutrements', 'buckle'), 15.954969),  (('accoutrements.', 'sun-set'), 15.954969),  (('accumulating', 'undirected'), 15.954969),  (('accuses', 'complains'), 15.954969),  (('adelaide', 'sidney'), 15.954969),  (('adornments', 'feastings'), 15.954969),  (('adulation', 'reward'), 15.954969),  (('adze', 'bolt'), 15.954969),  (('aeneas', 'odysseus'), 15.954969),  (('aesthetic', 'intellectual'), 15.954969),  (('affectations', 'lisps'), 15.954969),  (('agent', 'medium'), 15.954969),  (('agnus', 'dei'), 15.954969),  (('ague', 'convulse'), 15.954969),  (('aimedst', 'highly'), 15.954969),  (('alerted', 'heaven-clouds'), 15.954969),  (('alexandrian', 'pharos'), 15.954969),  (('algiers', 'tripoli'), 15.954969),  (('alimentive', 'amative'), 15.954969),  (('alleghanies', 'treating'), 15.954969),  (('altays', 'ghauts'), 15.954969),  (('amadis', 'tancred'), 15.954969),  (('amative', 'perceptive'), 15.954969),  (('amazon', 'paraguay'), 15.954969),  (('amazonia', 'patagonian'), 15.954969),  (('ambushes', 'opponents'), 15.954969),  (('amused', 'complacent'), 15.954969),  (('anatomist', 'phrenologist'), 15.954969),  (('andirons', 'straddle'), 15.954969),  (('angels', 'shirts'), 15.954969),  (('anon.', 'hand-mirror'), 15.954969),  (('antelope', 'burrowing'), 15.954969),  (('anticipated', 'cheaper'), 15.954969),  (('apathy', 'promptness'), 15.954969),  (('apex', 'apices'), 15.954969),  (('appoint', 'appointment'), 15.954969),  (('apprentice', 'churchman'), 15.954969) | (('abide', 'figures'), 13.508414),  (('acceptance', 'elegant'), 13.508414),  (('accounting', 'tastes'), 13.508414),  (('accustomed', 'usurpation'), 13.508414),  (('act', 'crawling'), 13.508414),  (('adjourn', 'immediate'), 13.508414),  (('adoption', 'energetic'), 13.508414),  (('affair', 'trusts'), 13.508414),  (('agony', 'terror'), 13.508414),  (('alarmed', 'proposal'), 13.508414),  (('ambition', 'distraction'), 13.508414),  (('ancient', 'modern'), 13.508414),  (('appearing', 'vanishing'), 13.508414),  (('arithmetic', 'ambition'), 13.508414),  (('attempts', 'applause'), 13.508414),  (('axis', 'axes'), 13.508414),  (('bag', 'tied'), 13.508414),  (('balanced', 'eel'), 13.508414),  (('barking', 'hoarsely'), 13.508414),  (('bathing', 'machines'), 13.508414),  (('bawled', 'murdering'), 13.508414),  (('beauti', 'ful'), 13.508414),  (('bed', 'various'), 13.508414),  (('belt', 'buttons'), 13.508414),  (('bitter', 'barley-sugar'), 13.508414),  (('bones', 'beak'), 13.508414),  (('bound', 'air.'), 13.508414),  (('brandy', 'choke'), 13.508414),  (('bright-eyed', 'terrier'), 13.508414),  (('brother', 'latin'), 13.508414),  (('burnt', 'eaten'), 13.508414),  (('buttered', 'toast'), 13.508414),  (('canvas', 'bag'), 13.508414),  (('carroll', '1865'), 13.508414),  (('chains', 'soldier'), 13.508414),  (('cherry-tart', 'custard'), 13.508414),  (('child-life', 'happy'), 13.508414),  (('childhood', 'gather'), 13.508414),  (('circle', 'exact'), 13.508414),  (('cleared', 'jelly-fish'), 13.508414),  (('condemn', 'death'), 13.508414),  (('counting', 'subjects'), 13.508414),  (('crocodile', 'improve'), 13.508414),  (('curving', 'graceful'), 13.508414),  (('custard', 'pine-apple'), 13.508414),  (('deepest', 'contempt'), 13.508414),  (('delay', 'cost'), 13.508414),  (('disappointment', 'empty'), 13.508414),  (('doubling', 'straightening'), 13.508414),  (('downward', 'antipathies'), 13.508414) |