

Worldreader Query Data Project

**A Capstone Project for the Data Science and Big Data course
at Universitat de Barcelona**



Team members:

Cary Lewis

Aina Pascual

Patricia Araguz

Enrique Rodríguez

Introduction & Project Scope

[Worldreader](#) is non-profit organization working to reduce illiteracy through its reading applications and sponsorship programs. The organization has a collection of over 40,000 books in more than 40 languages with the mission “to unlock the potential of millions of people through the use of digital books in places where access to reading material is very limited.”

55

COUNTRIES

464

SCHOOLS & LIBRARIES

692,190

MONTHLY READERS

6,139,558

READERS SINCE 2010

The Project Scope was to analyze queries made by users on the feature phone application by using clustering techniques to identify similar searches.

Worldreader provided our team with 6 CSV files consisting of over 3,000,000 queries and related information.

| customer, | country, | url, | query, | created_at |
|-----------|----------|--|---------------------|------------------------------|
| 157260, | "KE", | "/Search /Results?Query=New+Testament& Language=", | "New Testament", | "2016-12-27 15:48:16.893" |
| 157261, | "PH", | "/Search/Results?Query=circles", | "circles", | "2016-11-12 18:14:11.933" |
| 157261, | "PH", | "/Search /Results?Query=japanese", | "japanese", | "2016-11-18 17:15:54.19" |

Project Work

Data processing

Classification and Topic Modeling

P01

Data cleaning and language detection

- Data loading and general examination. Removing duplicates.
- Data cleaning: removal of punctuation, spaces before and after query, stop words, empty queries, queries of only numerical numbers and queries with special characters.
- Language detection.

P02

Language selection and error correction

- Select queries in English.
- Correct spelling errors.
- **P02A** Frequency of misspelled words.
- **P02B** Correction of misspelled words.

P03

Sampling & Descriptive Stats.

- Descriptive analysis of the sample.
- Graphical analysis of terms to identify possible themes: Word frequency, bigrams, and wordcloud.
- Sample selection of 20,000 queries run through the Google Books API supplementing the data.

P04

Classification

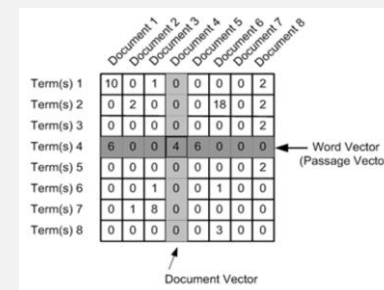
- Load the supplemented queries with the information pulled from the Google API.
- Segmentation of complete dataset through:
 - LDA Algorithm with best-match and with 5 results.
 - NMF Algorithm with best-match and 5 more results.
- Choose best fit model, test accuracy and stability.

Classification Algorithms : LDA vs NMF

LDA (Latent Dirichlet Allocation) and NMF (Non-negative Matrix Factorization) are unsupervised techniques for topic discovery in large document collections. Discover different topics that a set of documents represent and how much of each topic is present in a document (or corpus).

Each algorithm has a different mathematical underpinning:

- LDA is based on a bayesian probabilistic graphical modeling
- NMF relies on linear algebra.



| | Document 1 | Document 2 | Document 3 | Document 4 | Document 5 | Document 6 | Document 7 | Document 8 |
|-----------|------------|------------|------------|------------|------------|------------|------------|------------|
| Term(s) 1 | 10 | 0 | 1 | 0 | 0 | 0 | 0 | 2 |
| Term(s) 2 | 0 | 2 | 0 | 0 | 0 | 18 | 0 | 2 |
| Term(s) 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| Term(s) 4 | 6 | 0 | 0 | 4 | 6 | 0 | 0 | 0 |
| Term(s) 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| Term(s) 6 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| Term(s) 7 | 0 | 1 | 8 | 0 | 0 | 0 | 0 | 0 |
| Term(s) 8 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 |

Both algorithms take as input a bag of words matrix (i.e., each document represented as a row, with each columns containing the count of words in the corpus) and produce 2 smaller matrices:

- a document to topic matrix (no documents * k topics)
- a word to topic matrix (k topics * no words) that when multiplied together reproduce the bag of words matrix with the lowest error.

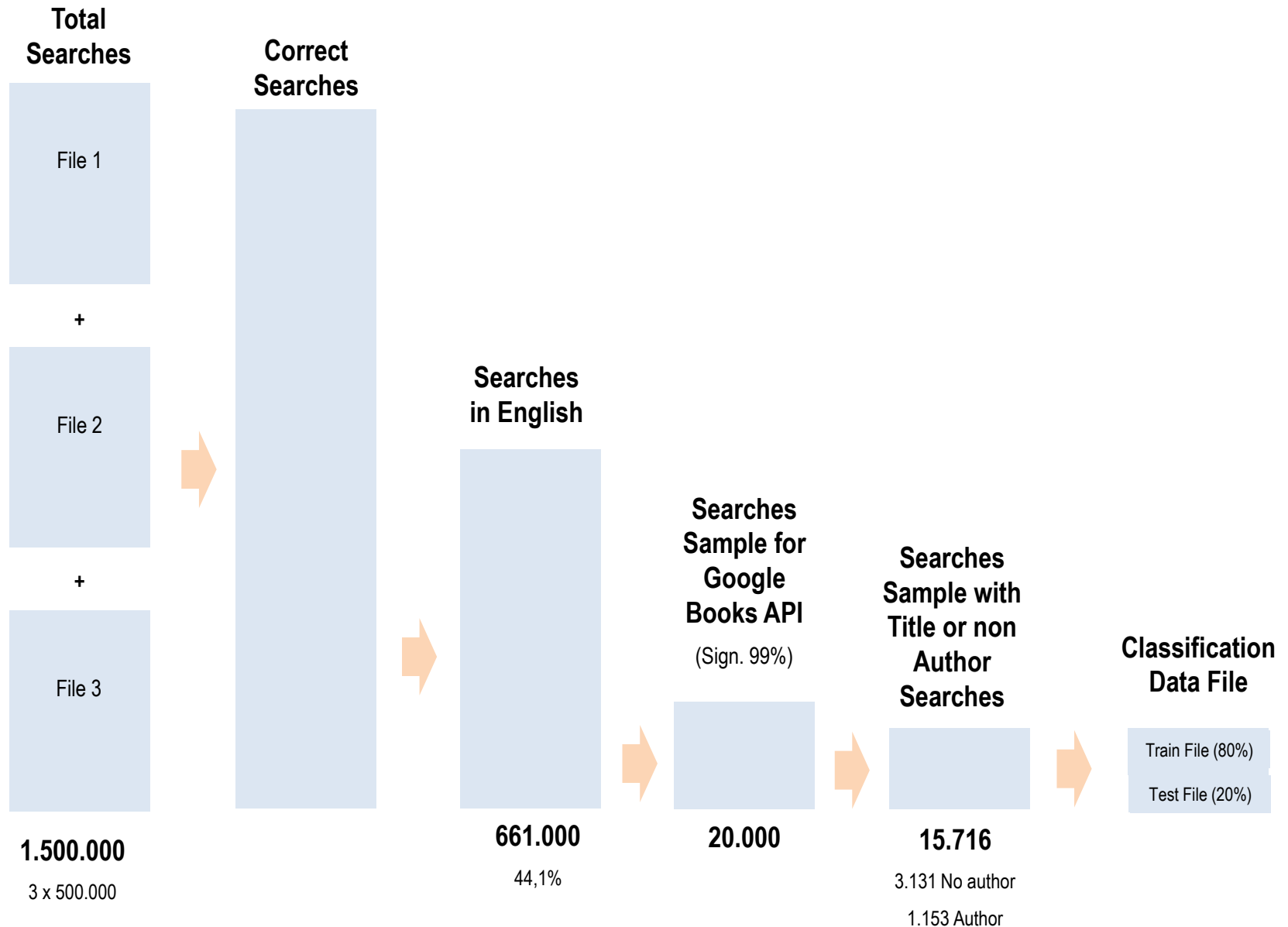
The output of the derived topics generated by both models involved assigning a numeric label to the topic and printing out the top words in a topic.

NMF and LDA are not able to automatically determine the number of topics and this must be specified.

NMF is usually [cheaper in computation](#) compared to LDA.

[NMF performs better](#) with shorter texts.

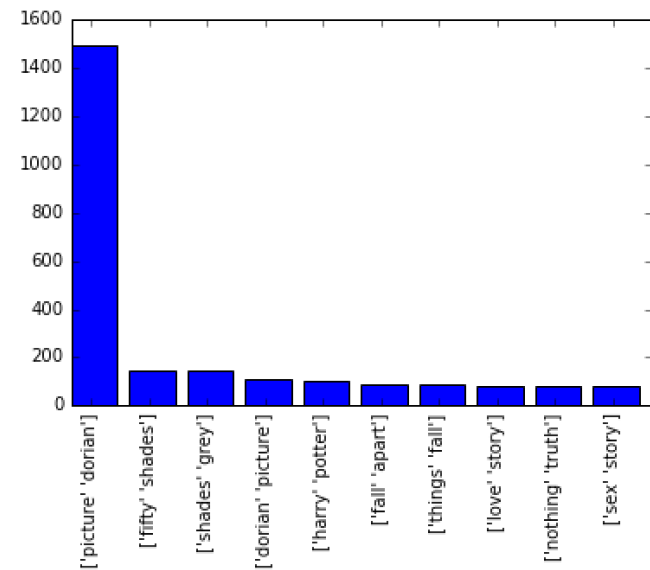
Searches file(s) process



Word cloud Sample

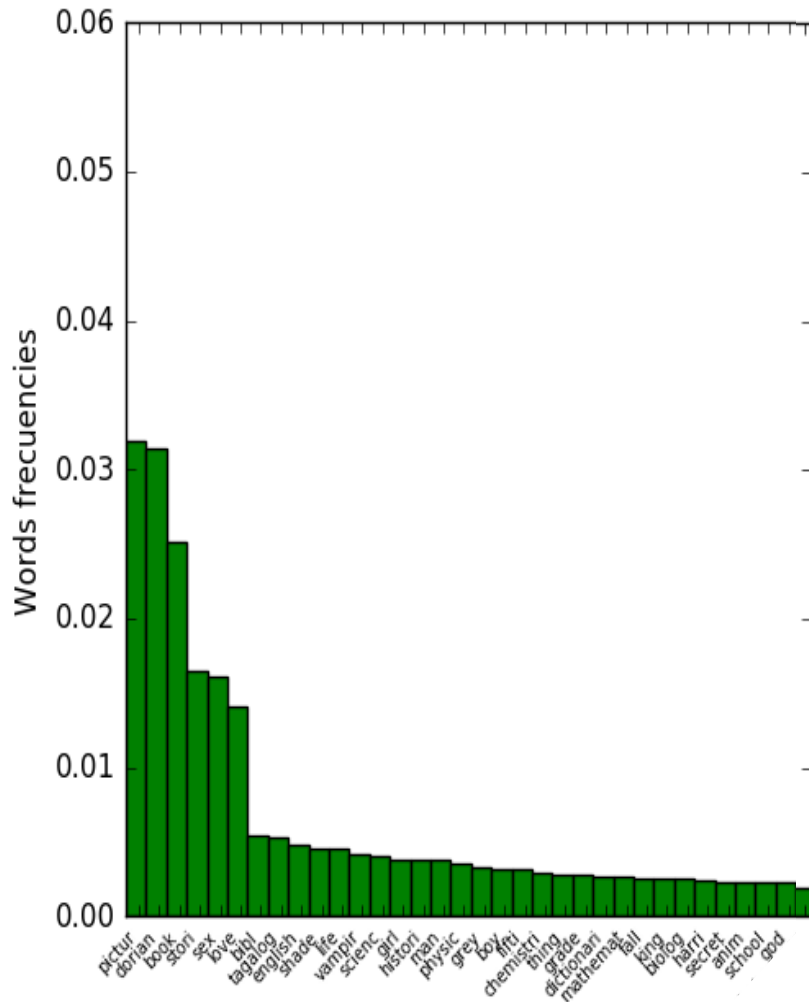


Bigrams Sample

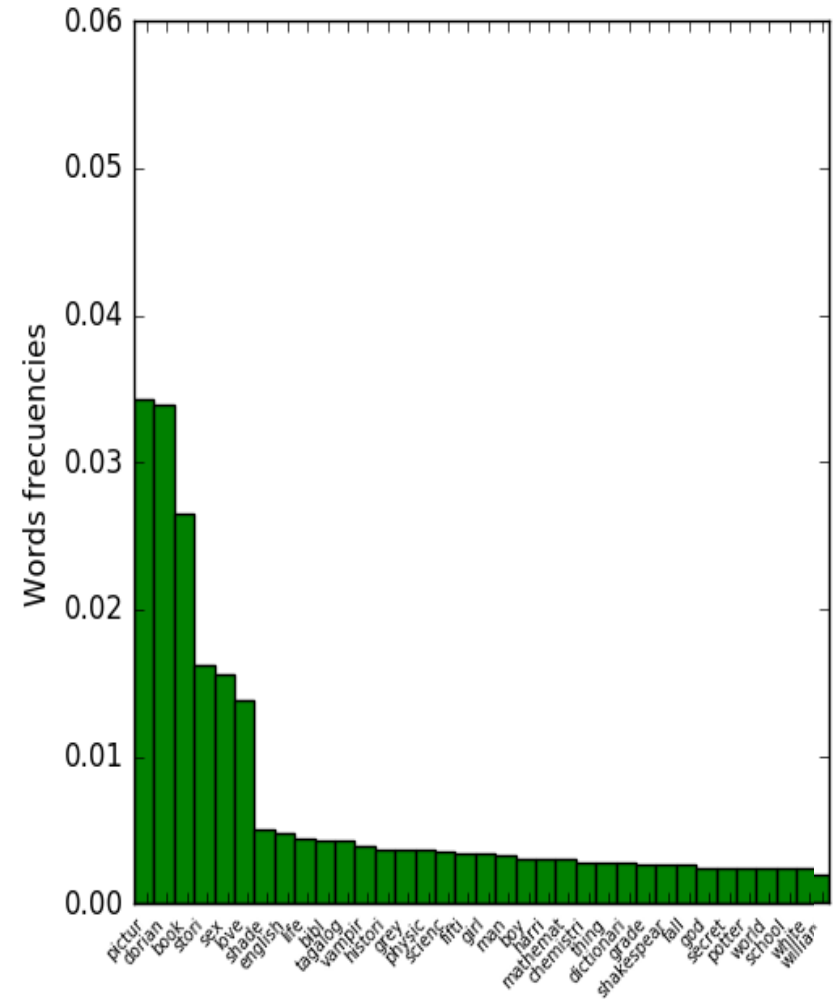


Sampling significance

Total Searches (N=661.000)



Sample (n=20.000)



Number of Topics

Iterative approach in both cases:

LDA

Perplexity is a commonly used measurement in information theory to evaluate how well a statistical model describes a dataset. Low perplexity indicates that the probability distribution of the model is good at predicting the sample.

| K | Perplexity |
|--------|---------------|
| K = 5 | 249.759001877 |
| K = 10 | 254.351936899 |
| K = 15 | 250.765157719 |
| K = 20 | 251.955603856 |
| K = 25 | 254.939880397 |

NMF

Tested with same K numbers. K = 15 was the one that performed best too.

LDA Topic classification : Complete Train vs. Best Match

LDA Complete Train (Time 37', words average 214)

| Topic 0: | Topic 1: | Topic 2: | Topic 3: | Topic 4: | Topic 5: | Topic 6: | Topic 7: | Topic 8: | Topic 9: | Topic 10: | Topic 11: | Topic 12: | Topic 13: | Topic 14: |
|----------|------------|----------|----------|-----------|----------|----------|----------|----------|----------|-----------|-----------|-----------|------------|-----------|
| student | english | harri | histori | sex | book | book | horror | dorian | book | comic | anim | shade | play | love |
| book | book | potter | world | sexual | manag | will | stori | wild | game | book | novel | fifti | shakespear | stori |
| learn | languag | lord | studi | book | system | vampir | film | gray | think | novel | thing | grey | poem | life |
| use | bibl | give | cultur | stori | use | find | fiction | pictur | will | music | fall | christian | tale | will |
| practic | word | will | book | work | research | new | ghost | stori | kid | time | african | book | romeo | can |
| mathemat | dictionari | war | polit | women | develop | seri | fan | portrait | get | life | apart | recip | juliet | book |
| includ | use | shall | new | psycholog | inform | power | the | beauti | can | work | stori | ana | othello | time |
| studi | includ | must | social | men | new | love | tale | edit | success | stori | africa | world | includ | live |

LDA Best Match (Time 37', words average 92)

| Topic 0: | Topic 1: | Topic 2: | Topic 3: | Topic 4: | Topic 5: | Topic 6: | Topic 7: | Topic 8: | Topic 9: | Topic 10: | Topic 11: | Topic 12: | Topic 13: | Topic 14: |
|----------|----------|----------|----------|----------|----------|------------|-----------|------------|----------|-----------|------------|-----------|-----------|-----------|
| book | book | will | histori | stori | love | shakespear | manag | law | dorian | book | sex | shade | book | dorian |
| student | languag | book | work | short | life | princ | busi | africa | gray | women | harri | fifti | histori | beauti |
| use | english | romanc | war | film | book | warrior | account | life | reader | sex | english | grey | scienc | portrait |
| includ | tagalog | seri | cultur | book | will | year | market | dictionary | pictur | food | sexual | christian | social | pictur |
| practic | stori | magic | new | horror | can | will | econom | legal | wild | vampir | dictionary | love | polit | begin |
| guid | play | ring | book | sex | live | play | develop | stori | young | men | potter | ana | human | depict |
| learn | student | king | art | adult | time | world | technolog | tale | first | journal | word | dark | studi | gray |
| provid | will | fantasi | poem | night | make | book | system | white | modern | recip | book | passion | new | physic |

NMF Topic classification : Complete Train vs. Best Match

NMF Complete Train (Time 15", words average 214)

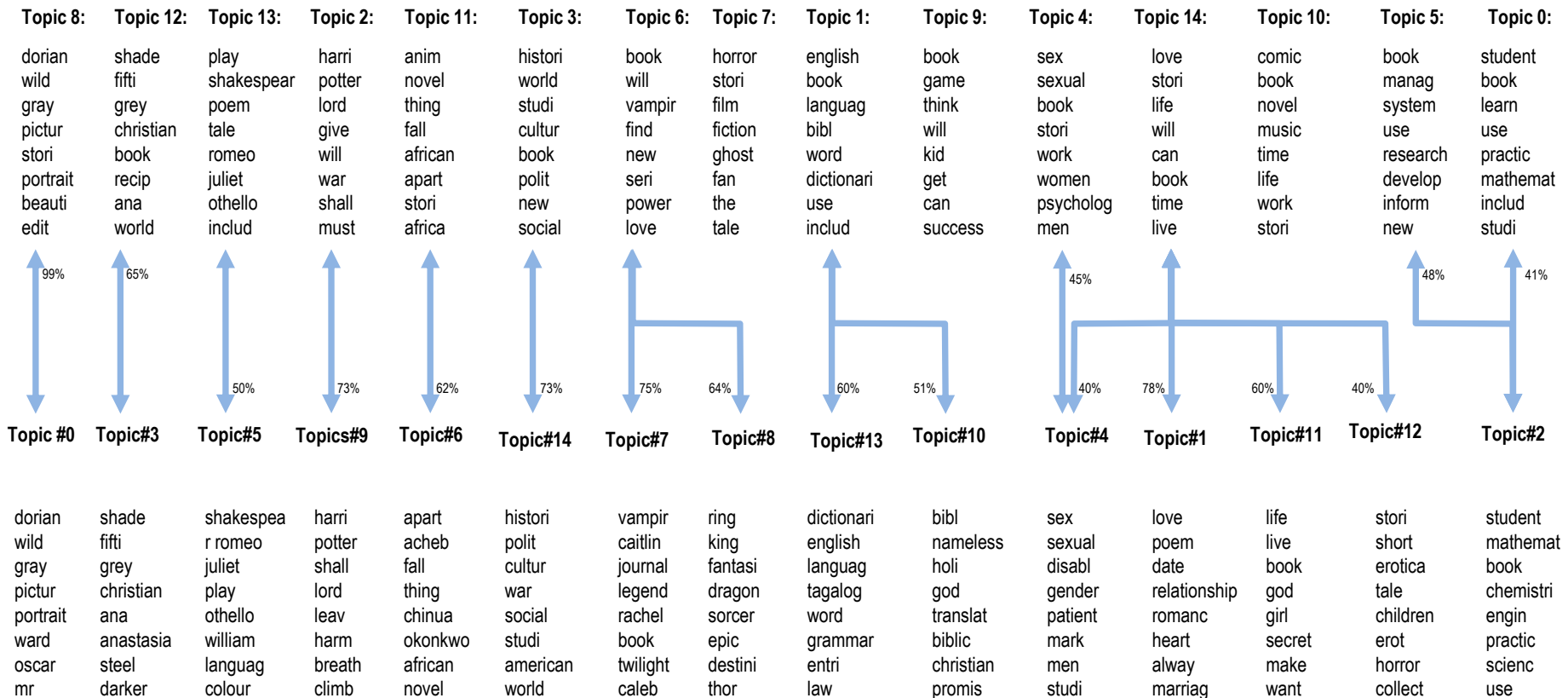
| Topic 0: | Topic 1: | Topic 2: | Topic 3: | Topic 4: | Topic 5: | Topic 6: | Topic 7: | Topic 8: | Topic 9: | Topic 10: | Topic 11: | Topic 12: | Topic 13: | Topic 14: |
|----------|--------------|-----------|-----------|----------|------------|----------|----------|----------|----------|-----------|-----------|-----------|------------|-----------|
| dorian | love | student | shade | sex | shakespear | apart | vampir | ring | harri | bibl | life | stori | dictionari | histori |
| wild | poem | mathemat | fifti | sexual | romeo | acheb | caitlin | king | potter | nameless | live | short | english | polit |
| gray | date | book | grey | disabl | juliet | fall | journal | fantasi | shall | holi | book | erotica | languag | cultur |
| pictur | relationship | chemistri | christian | gender | play | thing | legend | dragon | lord | god | god | tale | tagalog | war |
| portrait | romanc | engin | ana | patient | othello | chinua | rachel | sorcer | leav | translat | girl | children | word | social |
| ward | heart | practic | anastasia | mark | william | okonkwo | book | epic | harm | biblic | secret | erot | grammar | studi |
| oscar | alway | scienc | steel | men | languag | african | twilight | destini | breath | christian | make | horror | entri | american |
| mr | marriag | use | darker | studi | colour | novel | caleb | thor | climb | promis | want | collect | law | world |

NMF Best Match (Time 15", words average 92)

| Topic 0: | Topic 1: | Topic 2: | Topic 3: | Topic 4: | Topic 5: | Topic 6: | Topic 7: | Topic 8: | Topic 9: | Topic 10: | Topic 11: | Topic 12: | Topic 13: | Topic 14: |
|----------|----------|----------|-----------|----------|----------|-----------|----------|------------|----------|------------|-----------------|------------|--------------|-----------|
| dorian | life | dorian | sex | book | vampir | shade | bibl | english | harri | shakespear | histori | law | love | stori |
| gray | book | beauti | sexual | student | journal | fifti | nameless | languag | potter | romeo | cultur | legal | poem | short |
| wild | live | portrait | men | mathemat | book | grey | hero | dictionari | shall | play | polit | dictionari | marriag | erotica |
| pictur | time | pictur | male | scienc | twilight | christian | histori | word | lord | juliet | studi | entri | relationship | sexi |
| reader | world | depict | adult | engin | legend | ana | holi | oxford | leav | william | war | sourc | heal | night |
| remain | make | gray | enlighten | use | rachel | steel | heaven | tagalog | sky | othello | american sociat | crime | peter | erot |
| paint | god | begin | bodi | practic | diari | anastasia | way | grammar | drive | colour | centuri | term | richard | adult |
| portrait | famili | hedonist | character | studi | seri | darker | god | use | breath | classroom | | inform | romant | romanc |

LDA Topics classification : LDA Complete T. vs. NMF Complete T.

LDA Complete Train

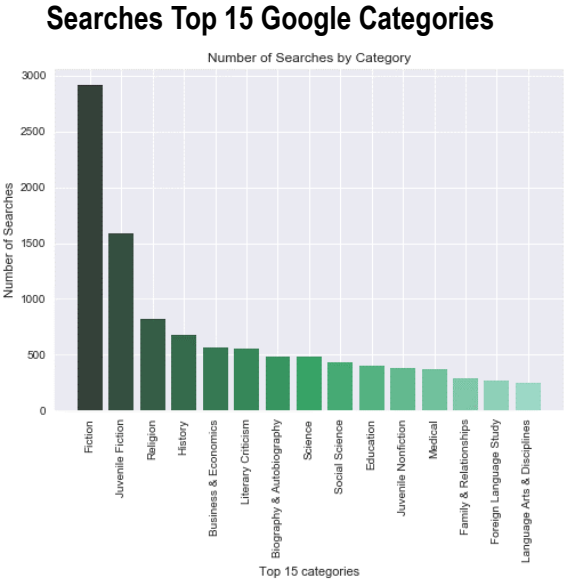
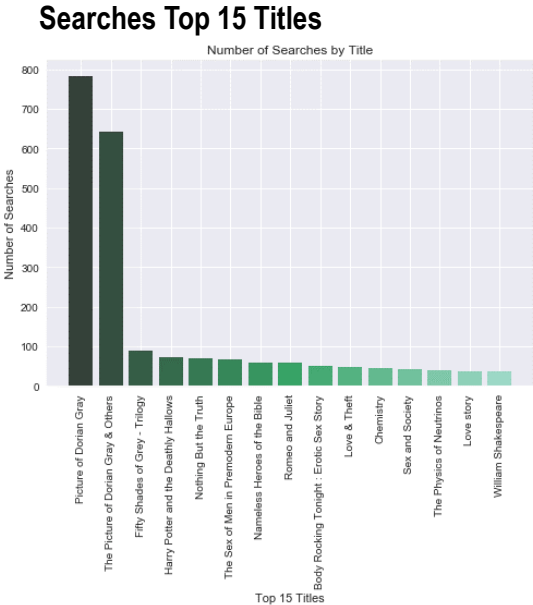


NMF Complete Train

NMF Complete Train Google Categories

| Novel | | | | | | | | | | | |
|---|--|--|---|--|---|---|---|--|--|--|-------------------|
| Topic #0 | Topic #5 | Topic #6 | Topic #7 | Topic #8 | Topic #9 | Topic #1 | Topic #11 | Topic #3 | Topic #12 | | |
| dorian wild gray pictur portrait ward oscar mr | shakespear romeo juliet play othello william languag colour | apart acheb fall thing chinua okonkwo african novel | vampir caitlin journal legend rachel book twilight caleb | ring king fantasi dragon sorcer epic destini thor | harri potter shall lord leav harm breath climb | love poem date relationship romanc heart alway marriag | life live book god girl secret make want | shade fifti grey christian ana anastasia steel darker | stori short erotica tale children erot horror collect | | |
| English Literature | | | African Liter. | | | Fiction/Juvenile Fiction | | | Family and relationships | | Erotic Literature |

| Book | | | | |
|--|---|---|--|--|
| Topic #4 | Topic #2 | Topic #14 | Topic #13 | Topic #1 |
| sex sexual disabl gender patient mark men studi | student mathemat book chemistri engin practic scienc use | histori polit cultur war social studi american world | dictionari english languag tagalog word grammar entri law | bibl nameless holi god translat biblic christian promis |
| Sex / Medical | Science | History | Language Study | Religion |



Google Books example categories

| Recurrent Titles | Google Category |
|------------------------|--------------------|
| Picture of Dorian Gray | Juvenile Fiction |
| Harry Potter | Juvenile Fiction |
| Fifty Shades | Fiction |
| Romeo and Julieta | Drama |
| Twilight | Literacy Criticism |
| Bible | Religion |
| Things Fall Apart | Social Science |

Insights for Worldreader and Recommendations

- Users use the open search field looking for specific titles or looking for literature related to a topic. Low searches with author names.
- High frequency titles searched need to be included if not in catalog: The Picture of Dorian Grey, The Bible, Fifty Shades of Gray, Romeo and Julieta, Things fall Apart, Twilight and Harry Potter Series.
- Solve user ID instability in order to understand user profiles, searches and behavior.
- Create a country by country strategy in order to enhance libraries.

Conclusions in terms of analysis

- LDA and NMF use information about the word co-occurrences to extract the latent topics of the data. For this reason, they promote the predictive capacity of the titles of the books in the categories.
- Topics generated by LDA are close to human understanding, in terms of grouping co-occurring words together. However, these topics may not necessarily be the ones that distinguish different groups of documents-sometimes enforcing the documents to be sparse and specific in topics may help.
- The results of LDA seem unstable and are different depending on the sample of 80% chosen.
- When perplexity was calculated LDA seemed to be more stable for 5 or 15 topics. 15 topics seemed to reflect more precisely the diverse range of documents.
- NMF can be mostly seen as a LDA of which the parameters have been fixed to enforce a sparse solution. It may not be as flexible as LDA if you want to find multiple topics in single documents (e.g., from long articles), but it usually works better with short texts of different nature.
- NMF is faster than LDA for short text analysis. Its computation time is lower.
- NMF seems to be a more stable model both with the best match and with all the suggestions.
- **For the purposes of this work, we think the results of NMF complete help us to understand better user's type of searches and recurrent topics.**

Possible next steps

- Classification of queries based off user information
 - Supplement data further with user profile information.
 - Supplement data further with information related to success or failure of the search results.
- Increase sample size to see if it improves the function of the models. Test different sample sets.
- Analysis of organization's catalog.
- Try to find other libraries with description. We could not access to any as good as Google Books.
- Compare the supplemented queries with the information extracted from the Google Books API against the organizational book catalog:
 - Identify the catalog books in high demand
 - A list of books in high demand and not part of the catalog
 - Create a recommenders system of the catalog based off the queries
- Query Classification Methodologies:
 - Test LDA removing words of low frequency.
 - Use PCA to try to find core topics within the data. Feature extraction method before NMF and LDA
 - Apply TextRank with the descriptions to extract keywords and run the models.
 - Apply technical analysis for graphs of the hyperonyms of the main search terms, using the NLTK library to retrieve WordNet hyperonyms.

Check out our project

GitHub and Jupyter Notebook: https://github.com/cnlewis/CAPE_stone

Blog: https://cnlewis.github.io/CAPE_stone

Thank you for your attention!

