Vivian Ellis

A Recommendation Engine for Relevant Segmentation Retrieval of Podcasts

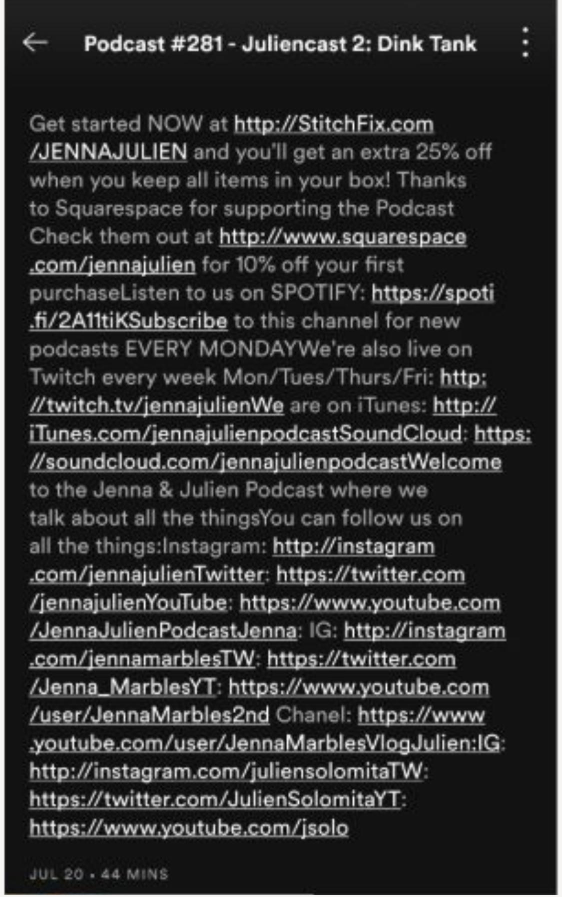
TABLE OF CONTENTS

01

INTRODUCTION

CURRENT SEARCH FUNCTIONALITY

- Combines episode title and description and matches a search query
- Episode title and description are written by the creators of the show

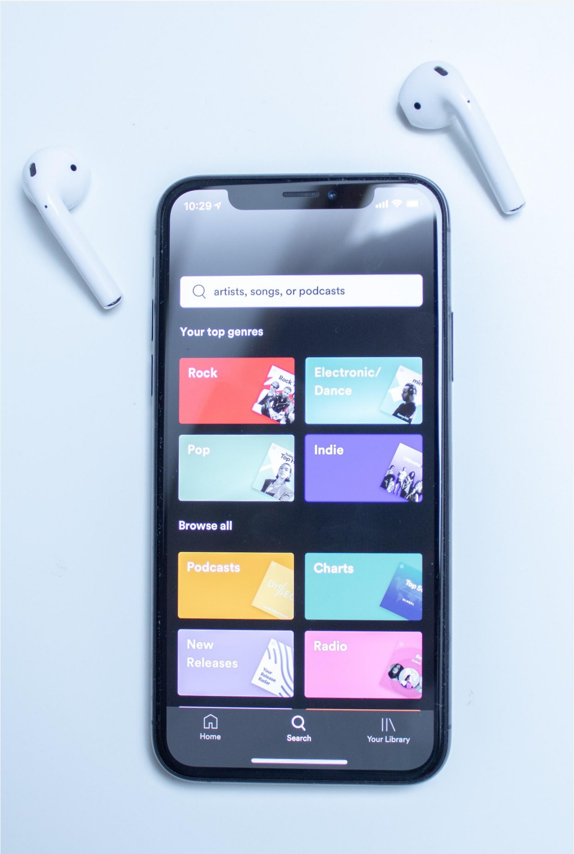


- A good description and title that accurately convey the content of the episode
- A bad description that contains ads and is not informative to a user

Recommendation Engine for Spotify

A scalable and experimental search engine is implemented in a three-phase model

- Query Expansion
- Topic Detection
- Retrieval and Ranking



02

RELATED WORKS

Sound and Music Recommendation with Knowledge Graphs

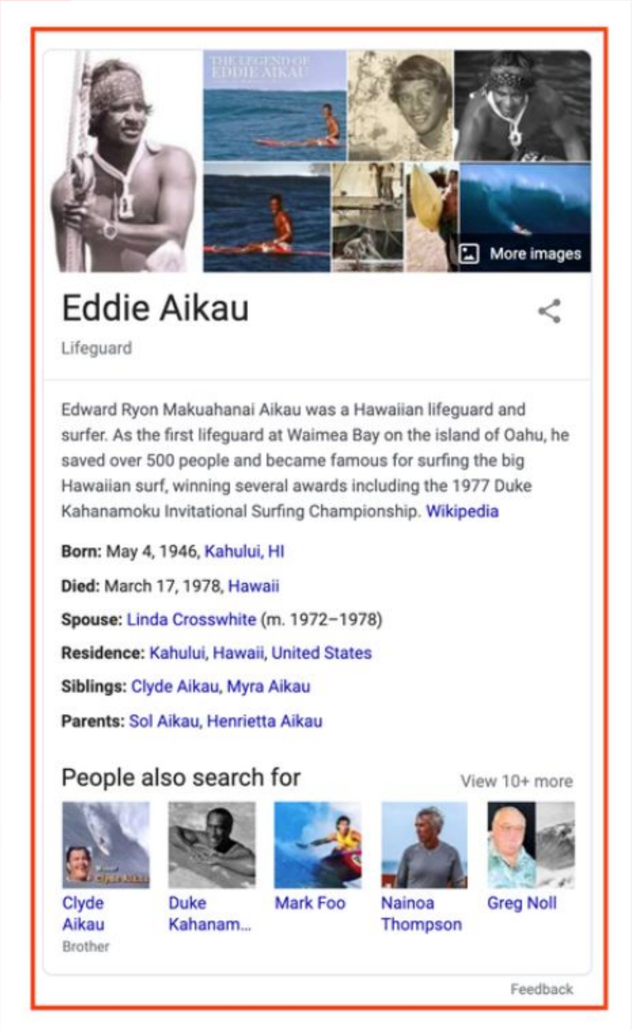
Oramas & Ostuni (2017)

- DBPedia is used to enrich information about named entities
- WordNet synset is used for the disambiguated words
- Entities are mapped together to create links between audio with the same tag
- User interaction

Introducing the Knowledge Graph: Things, Not Strings

Singhal (2012)

- Contains over 500 million objects and 35 billion facts from combined data sources:
    - Wikipedia
    - DBPedia
    - GeoNames
    - WordNet
    - CIA World Factbook



A Review Paper On The Application Of Knowledge Graph On Various Service Providing Platforms

Nigam & Paul (2020)

- Facebook creates a personalized view of content based upon location and friend connections
- Uber Eats recommends food based on menu items, location and cuisines
- Netflix increases viewership of TV and movies
- LinkedIn a knowledge graph for the professional world

TOPIC CLASSIFICATION IN SEARCH ENGINES

Latent Dirichlet Allocation for Tag Recommendation

Krestel & Fankhauser (2009)

CATEGORIZING WEB CONTENT

Tagging:
Organize, manage, and assist in the search for similar content

LATENT DIRICHLET ALLOCATION VS ASSOCIATION RULES

LDA:

- Implemented using Gibbs sampling
- Outperformed AR in recall, f-measure, and tf-idf

AR:

- Recommend generic tags and frequent tags
- Ineffective in tag recommendation

Latent Dirichlet Allocation in Web Spam Filtering.

Bíró & Szabó (2008)

- A novel multi-corpus LDA is integrated in a search engine to reduce spam
- F-measure of 46%

Topic-Aware Automatic Snippet Generation for Resolving Multiple Meaning on Web Search Result

Abe & Matsuhara (2018)

CLUSTERING WEB SEARCH RESULTS

- Set the number of topics to three
- Return 1,000 web pages for each query
- Text from each web page was extracted and used as input for the LDA
- F-measure calculated 95.6%

RELEVANCY AND RANKING IN INFORMATION RETRIEVAL

A Comparison of Information Retrieval Models

Pannu & James (2014)

EXPLORING ADVANTAGES AND DISADVANTAGES IN CLASSICAL IR TECHNIQUES

BM

Boolean Model

PRM

Probabilistic Retrieval Model

VSM

Vector Space Model

The documents and the queries needs to be represented in a way that allows mathematical operations to be performed

BM

- Based on set theory
- Requires dictionary of interesting words
- Query expressed using logical operators: AND, NOT, OR

- Does not return ranked list
- Exact matching criteria

PRM

- Determine the probability that a document within the corpus is relevant to the query
- Probability *p*: the number of relevant documents containing term over the total number of relevant documents
- Probability *q*: the number of irrelevant documents containing term over the total number of not relevant documents

- Probability *p* is set arbitrarily

VSM

- Represents documents using a vector of words
- Words have weights calculated by tf-idf
- Relevancy determined by cosine similarity between query vector and document vector

- Loss of recall and accuracy

Application of Topic Based Vector Space Model with WordNet

Wibowo & Handojo (2011)

EXPANDING VSM INPUT

- The query keyword becomes a list of related terms using WordNet
- Each term has an associated relation score
- Tests on 350 documents and expanded queries with a varying level of relation scores

03

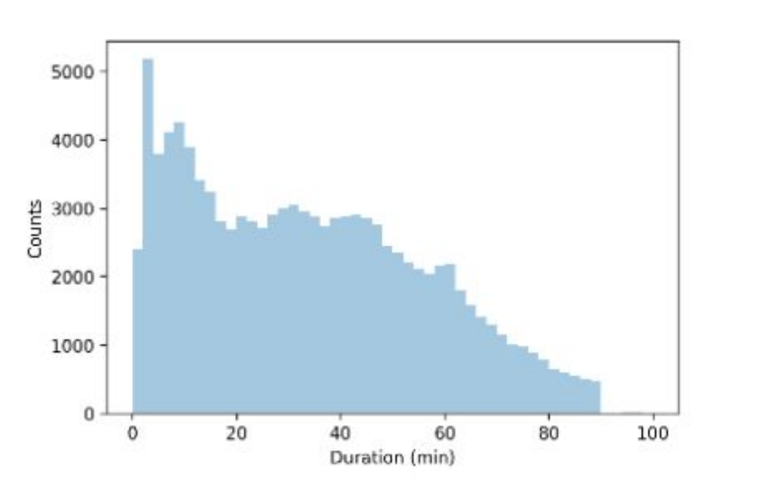
DATASET

ABOUT THE PODCASTS

- 100,000 text transcripts
- Contain only podcasts in English
- Noisy podcasts have been removed
- Unscripted to scripted and professionally developed to amateur
- Topics include arts & education, business & technology, comedy, educational, games, lifestyle & health, music, news & politics, society & culture, sports & recreation, stories, and true crime

DESCRIPTIVE STATISTICS

- Each episode ranges from less than a minute long to a maximum of 305 minutes
- An average length of 31.6 minutes
- Contains between 11 words and 43,504 words
- Average word count of 5,728



CONTENTS OF THE DATASET

- Transcripts – a JSON file
  - Broken up into 30-second intervals
  - Each word has the associated start and end time
  - Estimated the size of the transcript JSON file to be 12GB
- Metadata – a CSV file
  - Metadata about each episode: show uri, show name, show description, publisher, language, rss link, episode uri, episode name, episode description, and duration.
- Query – a JSON file
  - 50 queries

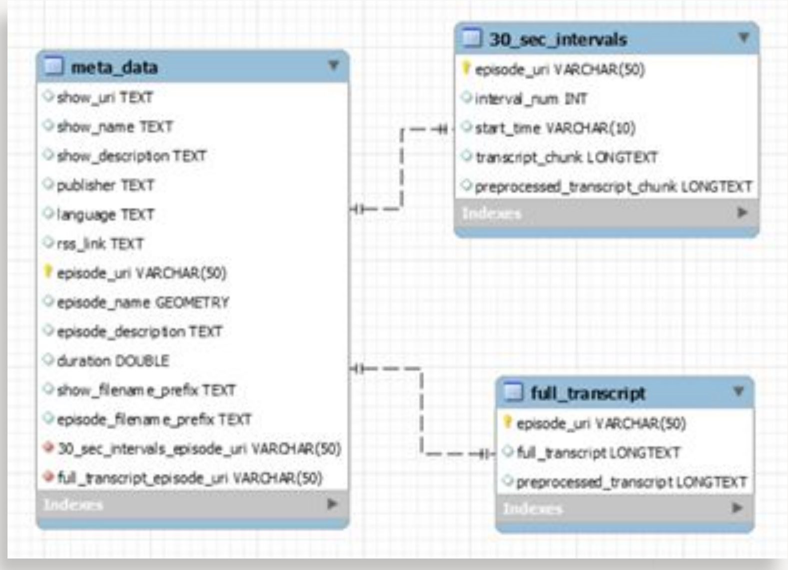
TRANSCRIPT JSON


```
 {"results":
 [{"alternatives":  // always only one alternative in these transcripts
   [{"transcript": "Hello, y'all, ... <30 s worth of text> ... ",
     "confidence": 0.8640950322151184,
     "words":  // list of words
     [{"startTime": "3s", "endTime": "3.300s", "word": "Hello,"},
...
]}]},
  {"alternatives": [
      {"transcript": "Aaron ... ",
       "confidence": 0.7733442187309265,
       "words": [
  {"startTime": "30s", "endTime": "30.200s", "word": "Aaron"}, ... ]}]},
  {"alternatives":  // last item in "results": a straight list of words with
"speakerTag"
   [{"words":
     [{"startTime": "3s", "endTime": "3.300s", "word": "Hello,", "speakerTag":
1},
      ...
{"startTime": "30s", "endTime": "30.200s", "word": "Aaron", "speakerTag": 1},
      ...

      {"startTime": "39.900s", "endTime": "40.500s", "word": "salon.",
"speakerTag": 2} ] }] }]
}
```

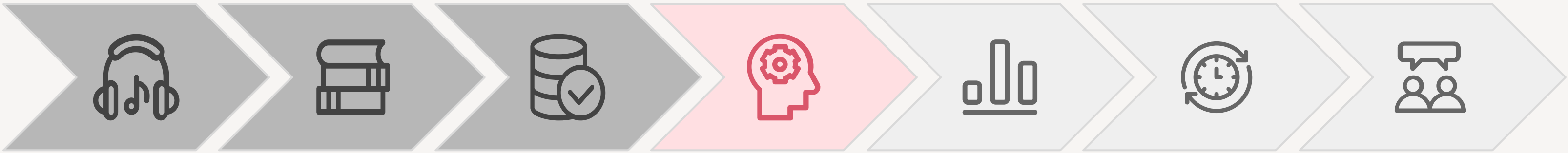
MYSQL DATABASE SCHEMA

- MySQL database is used to store metadata information and preprocessed transcripts
- JSON transcript files were parsed out to store the full transcript and each 30-second interval
- Full transcripts and 30-second intervals to undergo preprocessing before entering the database
- Episode uri is the primary key



04

METHODOLOGY

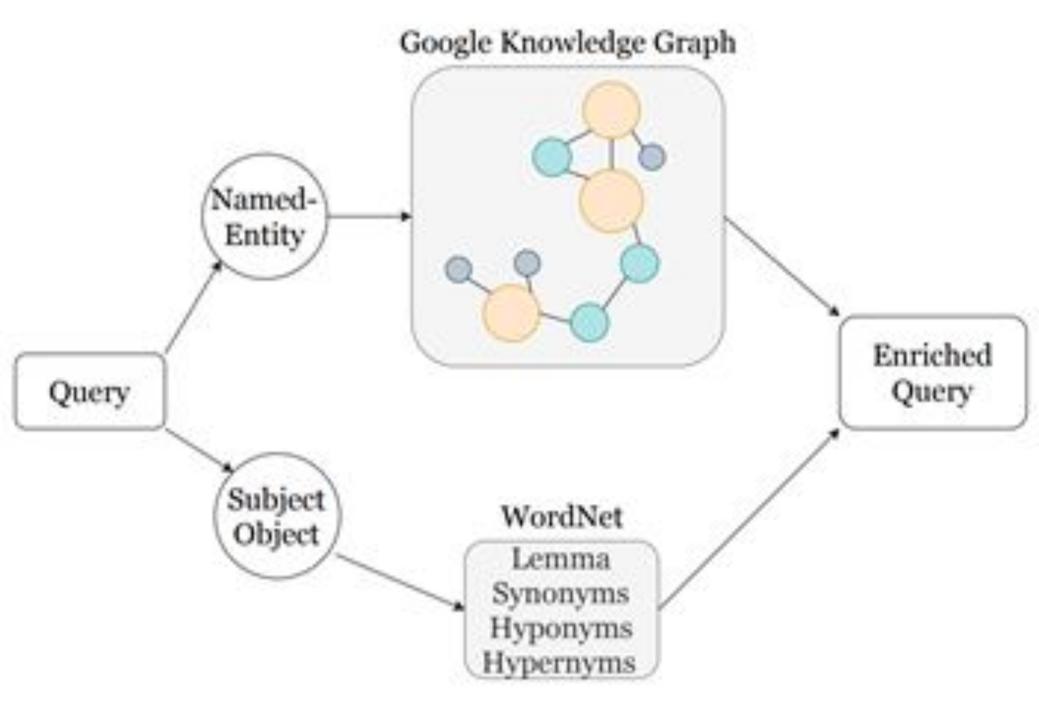


METHODOLOGY

- Query expansion
- Topic detection using LDA
    - Visualizing the LDA model
    - Coherence score
    - Tuning
    - Query tagging
- Retrieval and ranking using VSM

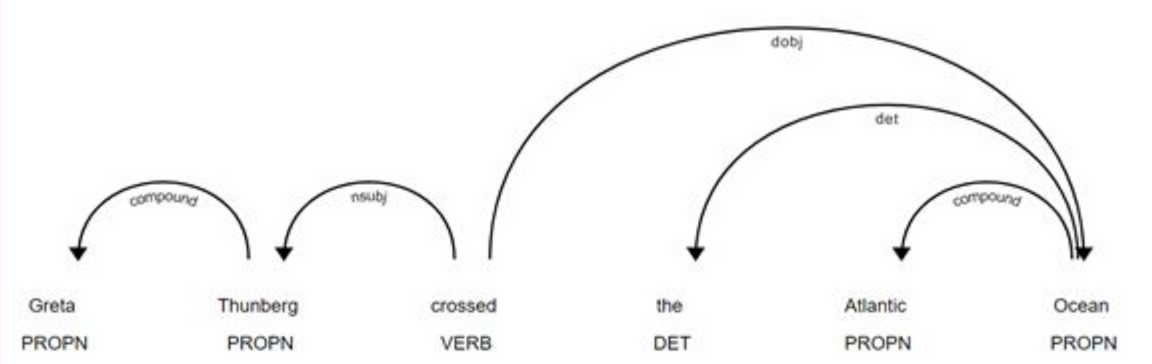
QUERY EXPANSION

Enriched query will uncover user semantics



COMPONENTS OF A SENTENCE

- The following sentence contains a direct object and a nominal subject, both of which are named entities that are linked to Google Knowledge Graph



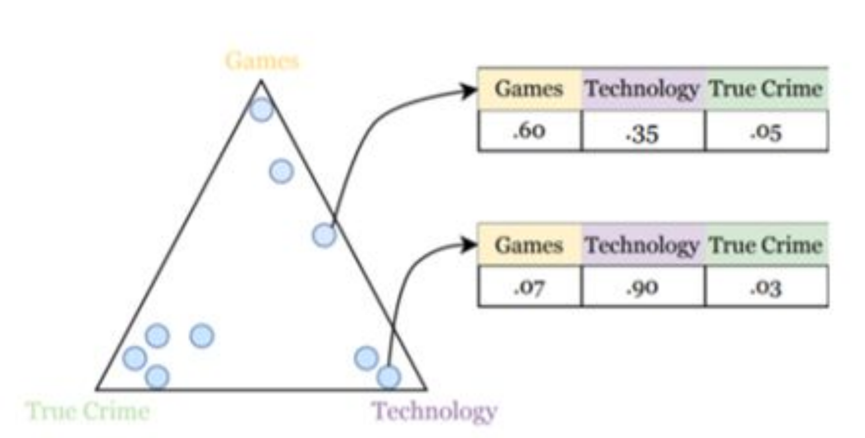
Text	Part-of-Speech	Dependency
How	ADV	advmod
was	AUX	ROOT
Greta	PROPN	compound
Thunberg	PROPN	poss
’s	PART	case
sailing	NOUN	compound
trip	NOUN	nsubj
across	ADP	prep
the	DET	Det
Atlantic	PROPN	Compound
Ocean	PROPN	Pobj
related	VERB	Acomp
to	ADP	Prep
global	ADJ	Amod
climate	NOUN	Compound
change	NOUN	Pobj

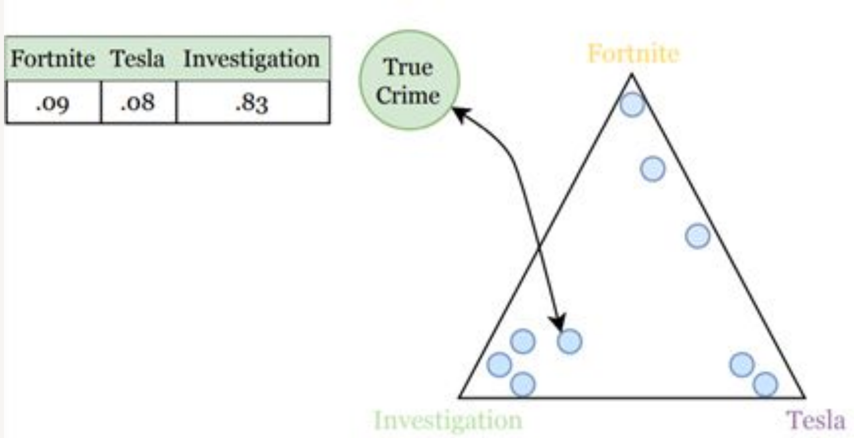
- Structure of the following sentence, “How was Greta Thunberg’s sailing trip across the Atlantic Ocean related to global climate change”?

INTERESTING WORDS

- Subjects, objects, and named entities
- Interesting words in the sentence “How was Greta Thunberg’s sailing trip across the Atlantic Ocean related to global climate change”:
  - Greta Thunberg, sailing trip, Atlantic Ocean, and climate change
- Interesting words that are not named entities are passed through WordNet
  - lemmas, synonyms, hyponyms, and hypernyms

LATENT DIRICHLET ALLOCATION

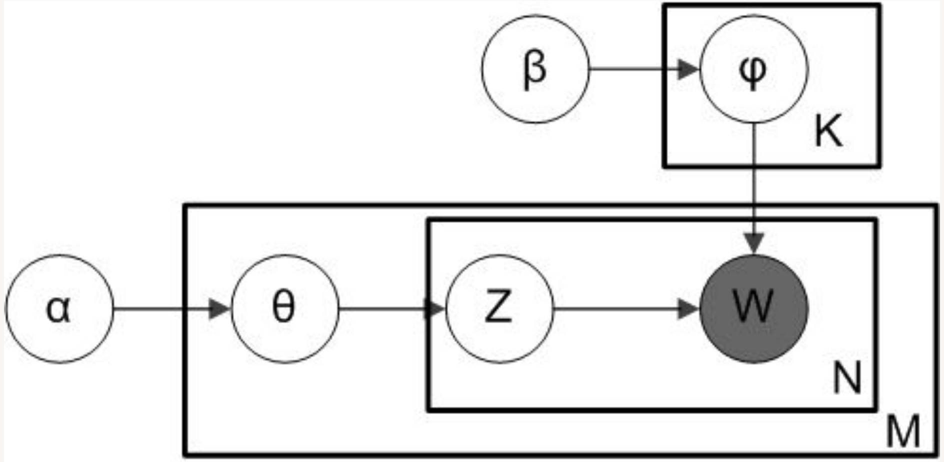




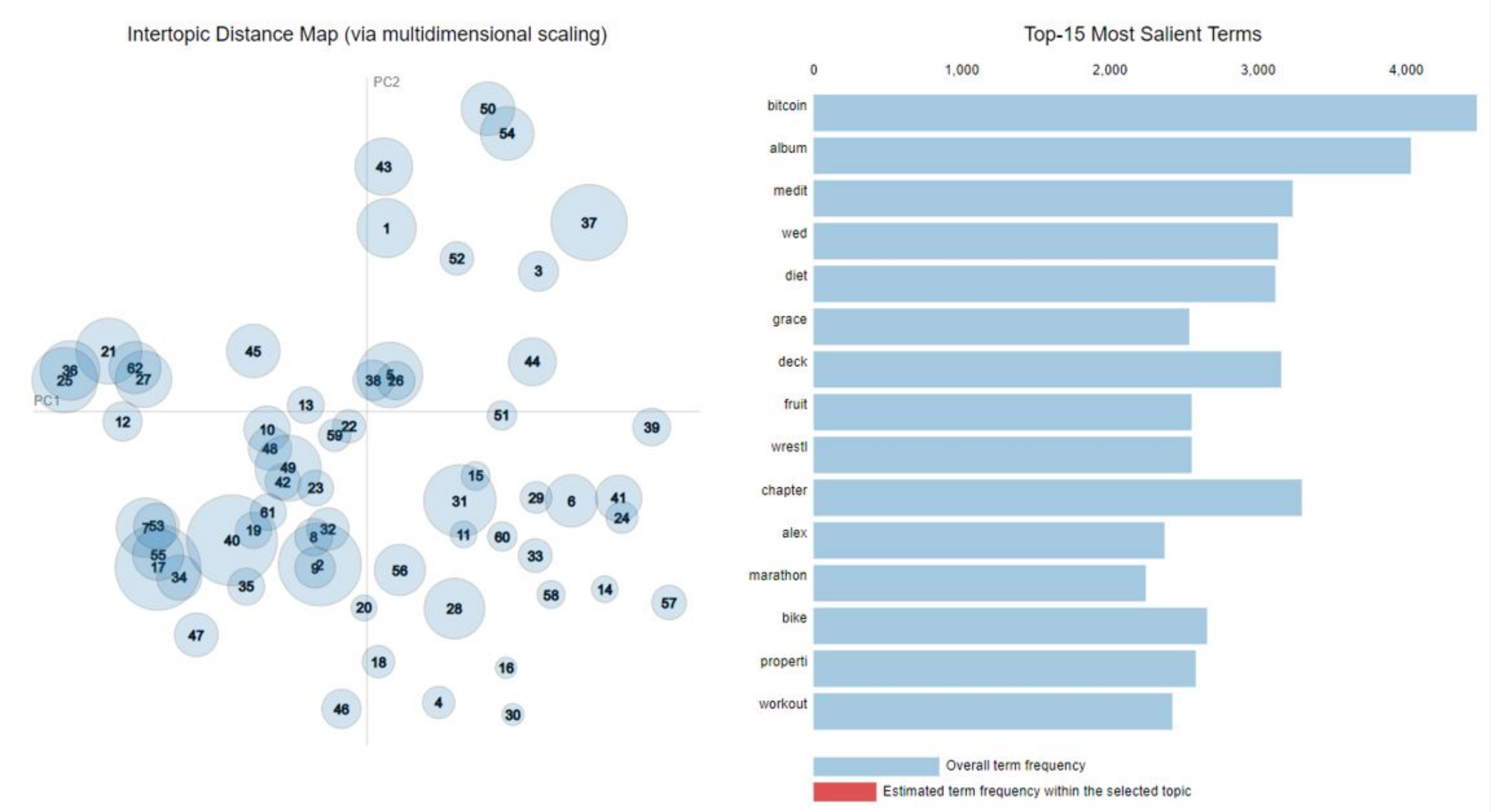
Per-Document Topics Distribution

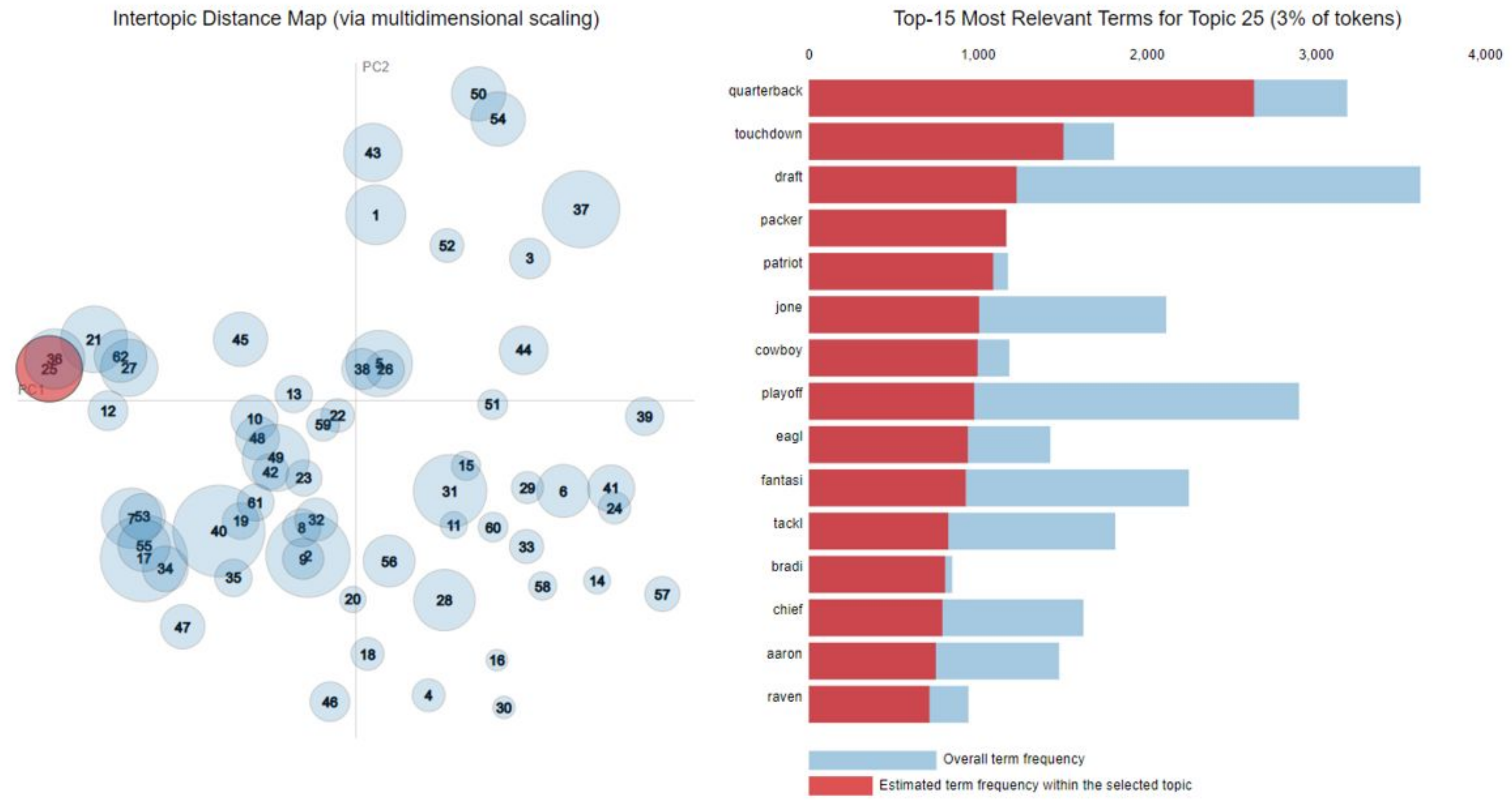
Per-Topic Word Distributions

- α is the Dirichlet prior on the per-document topic distributions
- β is the Dirichlet prior on the per-topic word distribution
- θ and ϕ, defined as the topic distribution for documents and the word distribution for topic *k* respectively
- *Z* is the list of topics
- M is the corpus
- N is a document



1. All the transcripts are tokenized to remove any punctuation
2. Boundary of each word is found to split the text into smaller units
3. Words that have fewer than 3 characters are removed
4. Each word is lemmatized to find a common root between words
5. Words are stemmed into root form to reduce inflection of words
6. A dictionary is created to tell us the number of times a word appears in the training set
7. The dictionary filters out very rare and very common words
8. A bag-of-words corpus is created


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Top-15 Most Salient Terms
0
1,000
2,000
3,000
4,000
bitcoin
album
medit
wed
diet
grace
deck
fruit
wrestl
chapter
alex
marathon
bike
properti
workout
Overall term frequency
Estimated term frequency within the selected topic



Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Top-15 Most Relevant Terms for Topic 25 (3% of tokens)
0
1,000
2,000
3,000
4,000
quarterback
touchdown
draft
packer
patriot
jone
cowboy
playoff
eagl
fantasi
tackl
bradi
chief
aaron
raven
Overall term frequency
Estimated term frequency within the selected topic

LDA COHERENCE SCORE

Coherence score reveals how similar the most relevant terms in a single topic are to each other to determine the interpretability of topics

The measure used to calculate coherence score is c_v and occurs in four steps:

1. Data segmentation
2. Probabilities of words
3. Confirmation measure
4. Mean of all confirmation measures is taken

LDA COHERENCE SCORE

The results are then normalized on a [-1,+1] scale where -1 indicates never occurs together, 0 for independence, and +1 for complete co-occurrence. The final coherence score of the LDA model is **0.537**

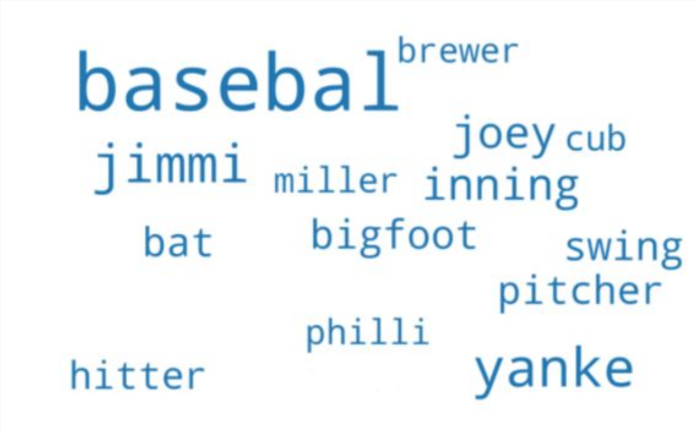
TUNING THE LDA MODEL

Beta	Alpha 0.01	0.31	0.61	0.91	symmetric	asymmetric
0.01	0.536	0.510	0.510	0.543	0.534	0.525
0.31	0.530	0.538	0.490	0.517	0.527	0.527
0.61	0.535	0.531	0.515	0.496	0.522	0.540
0.91	0.519	0.534	0.516	0.499	0.512	0.497
symmetric	0.502	0.520	**0.551**	0.524	0.509	0.510

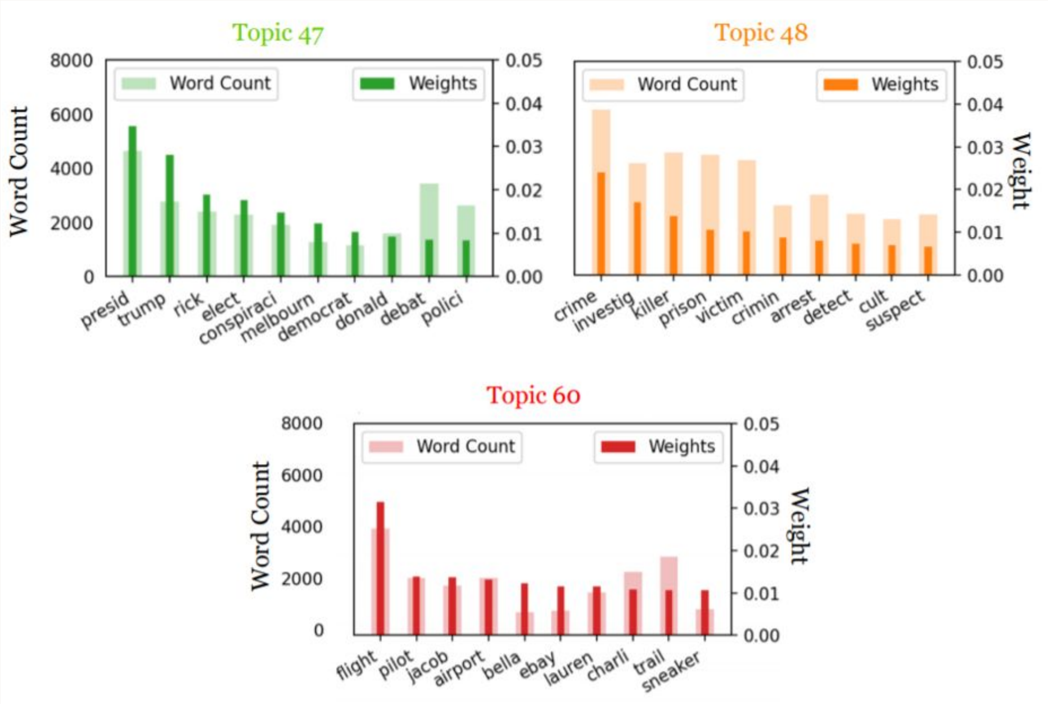
QUERY TAGGING

LDA THRESHOLD

- Let the minimum probability of the LDA model be **.25**
- Returning only topics with the highest scoring probabilities



- Let the query be ‘Babe Ruth’ the enriched GKG entity is returned and classified into Topic 12 with a **0.640** probability



"Who was involved in the assassination of JFK?"

- Interesting words are: involved, assassination and JFK

- Topic 47 with a 0.318 probability
- Topic 48 with a 0.176 probability
- Topic 60 with a 0.281 probability

IMPROVING GKG RESULTS

Relevance Score: 7,758.24 — 1 of 81

John F. Kennedy

35th U.S. President

Entity Types: Thing Person

John Fitzgerald Kennedy, often referred to by his initials JFK, was an American politician who served as the 35th president of the United States from January 1961 until his assassination in November 1963.

— Source: en.wikipedia.org (License)

View on Google

Image Source

Relevance Score: 2,438.61 — 2 of 81

John F. Kennedy International Airport

Airport in Queens, New York

Entity Types: Airport BusStation Thing Place

John F. Kennedy International Airport is an international airport in Queens, New York, USA, and one of the primary airports serving New York City.

— Source: en.wikipedia.org (License)

View on Google

Image Source

Relevance Score: 1,185.92 — 3 of 81

JFK

1991 film

Entity Types: Thing Movie

JFK is a 1991 American epic political thriller film directed by Oliver Stone. It examines the events leading to the assassination of United States President John F.

— Source: en.wikipedia.org (License)

View on Google

- Topic 47 with a 0.296 probability
- Topic 48 with a 0.586 probability

VECTOR SPACE MODEL

TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY (TF-IDF)

$V = (w_1, \dots w_n)$ where i is the word on the i^{th}- dimension

$\vec{q} = (x_1, \dots x_n)$ where x_i is the count of w_i in the query

$\vec{d} = (y_1, \dots y_n)$ where y_i is the count of w_i in the document

TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY (TF-IDF)

$$TF = \frac{Number\ of\ time\ the\ word\ occurs\ in\ the\ text}{Total\ number\ of\ words\ in\ text}$$

$$IDF = \frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ a\ word\ t\ in\ it}$$

LAYER ONE: EPISODE LEVEL

Query and corpus are represented as vectors with tf-idf, and return all episodes with a cosine similarity greater than **.1**

LAYER TWO: 30-SECOND INTERVALS

Query and the corpus of 30-second intervals are projected into a vector space. Cosine similarity is calculated with an arbitrary threshold of **.15**

SEGMENT ASSESSMENT

- Excellent (3): the segment conveys highly relevant information
- Good (2): the segment conveys highly-to-somewhat relevant information, is a good entry point for a human listener
- Fair (1): the segment conveys somewhat relevant information but is a sub-par entry point for a human listener
- Bad (0): the segment is not relevant

05

RESULTS

NORMALIZED DISCOUNTED CUMULATIVE GAIN

$$CG = \sum_{i=1}^{n} rel_i$$

$$DCG = \sum_{i}^{n} \frac{rel_i}{\log_2(i+1)}$$

$$iDCG = \sum_{i}^{n} \frac{idealrel_i}{\log_2(i+1)}$$

$$nDCG = \frac{DCG}{iDCG}$$

NORMALIZED DISCOUNTED CUMULATIVE GAIN

Query: What were people saying about the spread of the novel coronavirus NCOV-19 in Wuhan at the end of 2019?

rank	**segment**	*rel*
1	And the strain of coronavirus now are two different things. So yes coronavirus was there before but this coronavirus is of different strain. So is this coronavirus something which we can call the which can cause in World epidemic that is a question again, which is very very biased and very very subjective. I feel not bias. I'm so sorry for using that book. So why is objective is because...	2
2	Was reading through this because I was also like okay this coronavirus is something I should know about and I should think why is this really going to be an epidemic? So basically coronavirus has no vaccine that is the problem here and has no no proper medications which can directly which we can say that this medicine if you have coronavirus you can see it. So there are no proper precautionary...	2
3	Me give me the Deets bro. Give me the Deets the DJ. So this will happen details are in case you don't know what okay, so why should me the Deets bro? So the wahad virus is busy lately. K1 is a place in China scary. Yes, you'll excuse me. Before we continue. We don't want to spread panic. All right want to spread any like Panic. So before we continue we would like to tell you guys that I will pray...	1
4	Iris now before we start to everybody who's listening to this, thank you. Please don't listen to all my other episodes which will come every Thursday and I'm really glad that you guys are so let's do it. Now. What is Coronavirus coronavirus and its relation with the Corolla alcohol? There is no relation guys. So sorry coronavirus is basically when you look into the microscope, when you look at the coronavirus, you see that it is in the shape of...	2
5	Come over and listen, then the second one is moose meat. Oh, yeah, probably something that must be reported in Saudi Arabia in 2012. And then it's spread. Yeah. It's pretty little one but it wasn't there's no reason...	0

rank	**segment**	*rel*
6	Or coronavirus but in general, please wash your hands timely please maintain personal hygiene and hygiene when you cuff when you sneeze, please take a tissue paper and cover your mouth or your mouth when you cough and sneeze or take a cloth and do it wear masks. If you think you have a cold which can spread to others if it is a viral which is going viral if all these of all these precautions...	2
7	You already should yeah, I dropped the world the way yes, Lord cyclopaedia. Yeah, then I look down and it's like chemicals being spread your fucking donors back bending over or normal Le but honestly only 30 people of people are supposed...	0
8	What this virus and you show all the disease's we believe it because of this virus you should not have proper medications to cure yourself the Wuhan incident which we see from where the coronavirus outbreak has happened this the researchers which has been done on there and whatever from media channel channels and Outlets. We've heard that coronavirus stream is all this coronavirus stream...	1
9	Virus spread actually from from a seafood market in one is closely linked to the wanan seafood market where they sell this is fish my fish in animals do everything...	3

NORMALIZED DISCOUNTED CUMULATIVE GAIN

Query: What were people saying about the spread of the novel coronavirus NCOV-19 in Wuhan at the end of 2019?

DCG = 6.554
iDCG = 7.586
nDCG = ***.864***

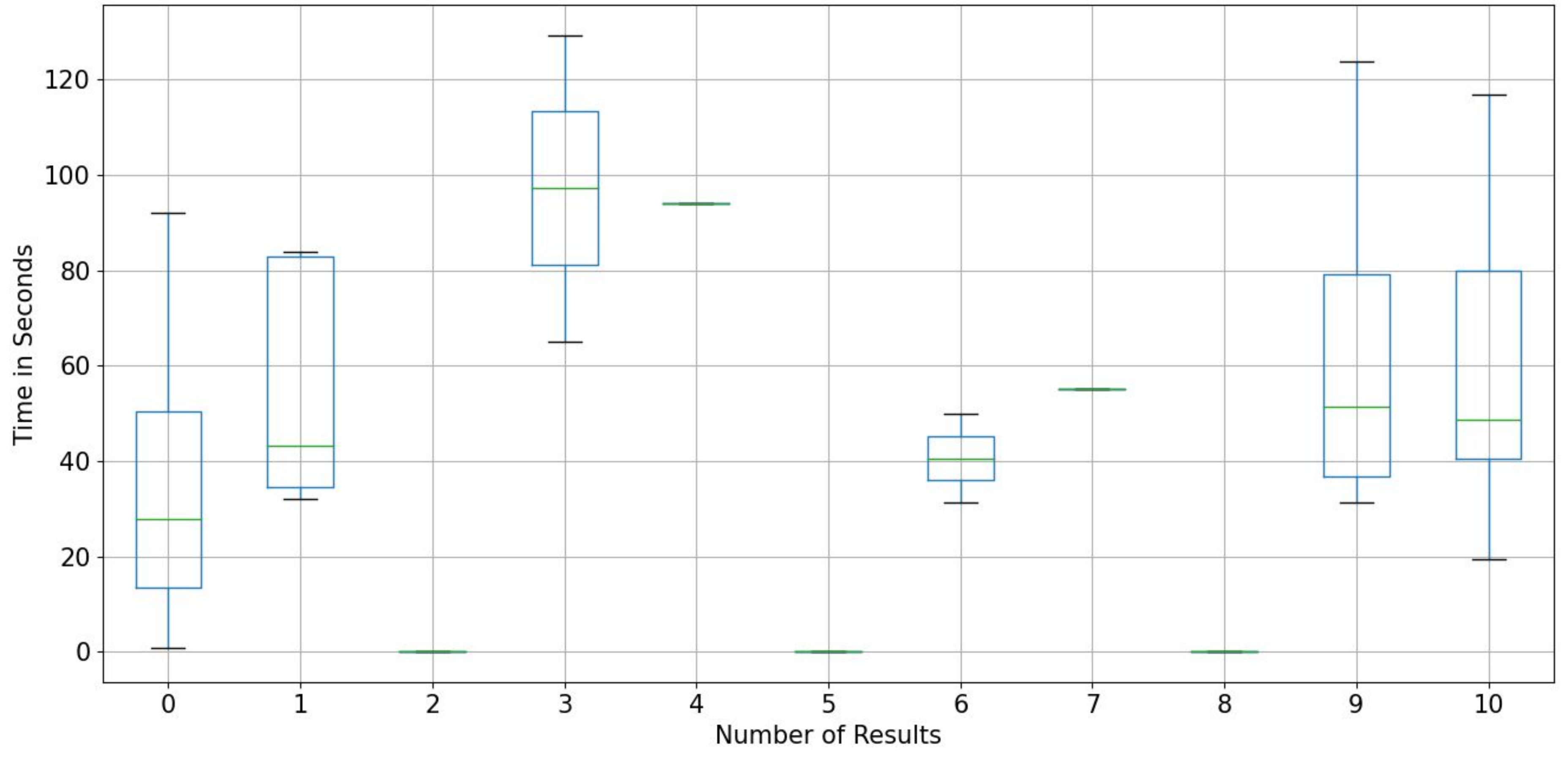
All nDCG values will be on the interval [0.0,1.0] where a nDCG of 1.0 is a perfect ranking

NORMALIZED DISCOUNTED CUMULATIVE GAIN

A test ran by lowering the threshold of the first layer VSM from .1 to .08 for all eight queries

query	$nDCG_{.1}$	$nDCG_{.08}$
1	0.864	0.789
2	N/A	N/A
3	0.897	0.897
4	0.557	0.557
5	N/A	0.6
6	1	1
7	N/A	1
8	0.567	0.737

AVERAGE TIME COST OVER 5 RUNS



06

FUTURE WORK

FUTURE WORK

- Incorporating user feedback, listening patterns, and content they follow for collaborative filtering
- Create a tailored experience for users to find relevant information based upon their listening habits on Spotify
- Expanding the LDA would allow the recommendation system to stay current
- Overcome VSM word mis-match
- DRMM to replace the second layer of VSM to increase the accuracy of the retrieved results

REFERENCES

Abe, H., Matsuhara, M., Chakraborty, G., & Mabuchi, H. (2018). Topic-Aware Automatic Snippet Generation for Resolving Multiple Meaning on Web Search Result. *2018 9th International Conference on Awareness Science and Technology (iCAST)*. doi:10.1109/icawst.2018.8517190

Bíró, I., Szabó, J., & Benczúr, A. A. (2008). Latent dirichlet allocation in web spam filtering. *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web - AIRWeb '08*. doi:10.1145/1451983.1451991

Krestel, R., Fankhauser, P., & Nejdl, W. (2009). Latent Dirichlet Allocation for Tag Recommendation. *Proceedings of the Third ACM Conference on Recommender Systems - RecSys '09*. doi:10.1145/1639714.1639726

Nigam, V. V., Paul, S., Agrawal, A. P., & Bansal, R. (2020). A Review Paper On The Application Of Knowledge Graph On Various Service Providing Platforms. *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. doi:10.1109/confluence47617.2020.9058298

Oramas, S., Ostuni, V. C., Noia, T. D., Serra, X., & Sciascio, E. D. (2017). Sound and Music Recommendation with Knowledge Graphs. *ACM Transactions on Intelligent Systems and Technology, 8*(2), 1-21. doi:10.1145/2926718

Pannu, M., James, A., & Bird, R. (2014). A Comparison of Information Retrieval Models. *Proceedings of the Western Canadian Conference on Computing Education - WCCCE '14*. doi:10.1145/2597959.2597978

REFERENCES

Singhal, A. (2012, May 16). Introducing the Knowledge Graph: Things, not strings. Retrieved July 20, 2020, from https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

Sullivan, D. (2020, May 20). A reintroduction to our Knowledge Graph and knowledge panels. Retrieved July 20, 2020, from https://blog.google/products/search/about-knowledge-graph-and-knowledge-panels/?_ga=2.87765016.1275668186.1595206893-719792272.1595206893

Wibowo, A., Handojo, A., & Halim, A. (2011). Application of Topic Based Vector Space Model with WordNet. *2011 International Conference on Uncertainty Reasoning and Knowledge Engineering*. doi:10.1109/urke.2011.6007864

Zipf's Law: Modeling the Distribution of Terms. (2009, July 04). Retrieved July 21, 2020, from https://nlp.stanford.edu/IR-book/html/htmledition/zipfs-law-modeling-the-distribution-of-terms-1.html

07

DISCUSSION

Thank you for your time
Questions, comments, and feedback is now welcome