Semantico

SEMANTICO: SMART WIKI SEARCH

BY PATRICK YAMIN

About Me:

- From Los Angeles, CA
- Data Scientist and aspiring ML Engineer
- Fascinated with **search**, one of the most useful tools we utilize on daily basis
- Motivations for capstone were inspired by Stack Overflow search algorithms, Google, and new SOTA web search like “Perplexity”

Business Problem - Why Semantic Search?

SEMANTIC VS KEYWORD SEARCH

- SEMANTIC SEARCH REPRESENTS AN ADVANCED APPROACH TO SEARCH QUERIES THAT FOCUSES ON UNDERSTANDING THE CONTEXT AND INTENT BEHIND A USER'S QUERY, RATHER THAN RELYING SOLELY ON MATCHING KEYWORDS.

- SEMANTIC SEARCH CAN INTERPRET AND ANALYZE THE NUANCES AND MEANINGS OF WORDS WITHIN THE CONTEXT, OFFERING MORE ACCURATE AND RELEVANT RESULTS.

DATA MINING & UNDERSTANDING

WIKIPEDIA

- Collected through a series of API calls through the wikipedia-api

- Articles that were extracted were “Featured articles” - 6,428 articles

- Best articles Wikipedia has to offer, as determined by wiki editors for accuracy, neutrality, completeness, and style

DATA CLEANING & PREPROCESSING

- Lower-casing/Standardizing text
- Removing Special Characters and Punctuation
- Replacing Line Breaks

- Stop Word Removal - Eliminating common words that add little value to the analysis, such as "the", "is", etc., using NLTK's predefined list of stop words.
- Lemmatizing
- Tokenization

- From here we can create our word embeddings

CHALLENGES & OBSTACLES

TOPIC MODELING

- LDA - Latent Dirichlet Allocation
- NMF - Non-Negative matrix factorization

Topics from NMF

```
Topic 1: ['state', 'would', 'new', 'president', 'government', 'party', 'american', 'united', 'war', 'election']
Topic 2: ['game', 'player', 'team', 'character', 'series', 'released', 'one', 'also', 'sonic', 'version']
Topic 3: ['film', 'character', 'role', 'time', 'one', 'scene', 'said', 'also', 'best', 'series']
Topic 4: ['specie', 'also', 'may', 'found', 'one', 'bird', 'used', 'known', 'year', 'two']
Topic 5: ['ship', 'fleet', 'gun', 'two', 'british', 'navy', 'battleship', 'sea', 'naval', 'squadron']
Topic 6: ['first', 'season', 'team', 'match', 'league', 'club', 'second', 'two', 'player', 'run']
Topic 7: ['album', 'song', music', 'band', 'released', 'single', 'first', 'record', 'one', 'number']
Topic 8: ['city', 'river', 'park', 'area', 'state', road', 'highway', 'north', 'building', 'new']
Topic 9: ['king', 'work', 'henry', 'year', 'first', 'one', 'also', 'time', 'england', 'later']
Topic 10: ['army', 'force', 'division', 'war', 'german', 'attack', 'battle', 'japanese', 'troop', 'battalion']
```

UNSUPERVISED LEARNING

- K-Means Clustering
- PCA

ANNOY

- Approximate Nearest Neighbors Oh Yeah
- Mixed Results

SENTENCE TRANSFORMERS

- Best Performing Model
- Based on cosine similarity score

SENTENCE TRANSFORMERS & MULTI-VECTOR SEARCH

- State-of-the-art sentence, text and image embeddings
- For search or information retrieval, we would like to see accurate search results, encompassing wide-range queries
- SentenceTransformers is built on PyTorch and Transformers architecture

Best Model: all-MiniLM-L6-v2

Retrieve & Re-Rank Pipeline

- In order to improve our semantic search, cross-encoders and bi-encoder models were implemented in a search pipeline
- Retrieval system might retrieve documents that are not that relevant for our search query
- Use Cross-Encoder for re-ranking, scores the relevancy of all documents for the given search query

Search with Retrieval & Re-ranker

Search Query: “coldest place earth”

```
Top-3 Cross-Encoder Re-ranker hits
        0.571   Antarctica antarctica earth southernmost leastpopulated
continent situated almost entirely south antarctic circle surrounded southern
ocean also known antarctic ocean contains geographic south pole antarctica
fifthlargest continent 40 larger europe area 14200000 km2 5500000 sq mi
antarctica covered antarctic ice sheet average thickness 19 km 12 mi antarctica
average coldest driest windiest continent highest average elevation mainly polar
        -3.977  Neptune neptune eighth farthest planet sun fourthlargest planet
solar system diameter thirdmostmassive planet densest giant planet 17 time mass
earth slightly massive neartwin uranus neptune denser physically smaller uranus
       -4.396  Farthest South farthest south refers southerly latitude reached
explorer first successful expedition south pole 1911 significant step road pole
discovery land south cape horn 1619 captain james cook crossing antarctic circle
1773 earliest confirmed sighting antarctic mainland 1820 late 19th century
```

Coldest place

Add lang

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

Coldest place may refer to:

- The coldest place on earth, the Pole of Cold in Antarctica.
- The coldest temperatures seen in the known universe. See Absolute zero § Very low temperatures.

CONCLUSIONS AND STRETCH GOALS

Reranking can be used in a RAG pipeline

• Sentence Transformers are SOTA • Search is not yet solved but vastly improving	• Utilize decoder models, like GPTs to summarize text from relevant documents • Manually annotate/label data	• RAG • Retrieval Augmented Generation • Pipeline chained with LMs

THANK YOU

GitHub

www.github.com/pyamin1878

substack

https://substack.com/@patrickyamin

Patrick Yamin