Amazon Reviews Topic Modelling

What is LDA?

- LDA stands for Latent Dirichlet Allocation.
- In NLP the use for an LDA model is to help with uncover hidden themes or topics within a collection of documents or texts
- LDA is a probabilistic model that assumes each document is a mixture of various topics and each word in the document is attributable to one of these topics

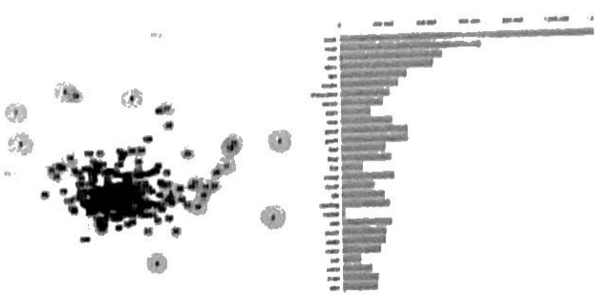
LDA Methods

- LDA is a topic model
- Given the input of data (summary and reviews) it will:
  - Tokenize the sentences to words and clean the data using stopwords
  - Then it makes a dictionary full of words
  - Then it calls gensim which are initially assigned randomly to words in documents. This process continues until the model converges, providing a meaningful representation of topics within the given data.
- Uses CPUs as workers in order to help produce the model

LDA Analysis/Results

- For the LDA model there was a lot of steps in order to get the data in the right format for the model.
- It is very CPU intensive had to ensure it used 20 cpu's with 24gb for an efficient run time.
- The model would crash due to not enough memory especially when trying to run the 55 million dataset for books.
- There is no ideal topic decides that is parameter that the user has to tune per model

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



What is the Problem?

- We found a large dataset of Amazon reviews organized into a variety of categories, and wanted to explore the nature of the data
- We decided to use topic modeling approaches to learn about the structure of the topics in the datasets
- We used two models with different approaches to topic modeling, LDA and BERTopic, and trained them on various categories of the dataset


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "vote": 5,
  "style": {
    "Format:": "Hardcover"
  },
  "reviewText": "I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. The music is at times hard to read because we think the book was published for singing from more than playing from. Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


- reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- asin - ID of the product, e.g. 0000013714
- reviewerName - name of the reviewer
- vote - helpful votes of the review
- style - a disctionary of the product metadata, e.g., "Format" is "Hardcover"
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)
- image - images that users post after they have received the product

https://github.com/hamidc10/NLP_FP

What is BERTopic?

- BERTopic is a topic modeling model based on BERT
- It is an unsupervised model, and has a modular pipeline which can preprocess data, generate embeddings, and cluster them into topics.

BERTopic Methods

- Given input documents, this model will:
  - Tokenize and clean input (CountVectorizer)
  - Generate sentence embeddings (SBERT)
  - Reduce dimensionality of embeddings (UMAP)
  - Cluster reduced embeddings (hDBSCAN)
  - Generate topics for each cluster (c-TF-IDF)
- Used GPU-accelerated embedding generation, UMAP, and hDBSCAN to speed up training

BERTopic Analysis/Results

- I liked the modular nature of the BERTopic pipeline
  - There was not a lot of preprocessing
  - It was easy to tweak the parameters at each step
- Precalculating embeddings and using GPUs greatly sped up the process
- It was difficult to train very large datasets, sometimes the kernel would die