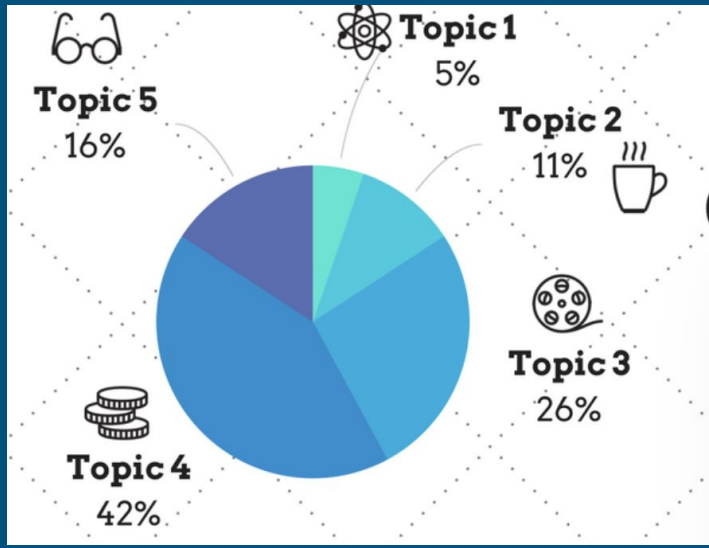
Topic modelling (Inflation Tweets)

Disha Patel
B17

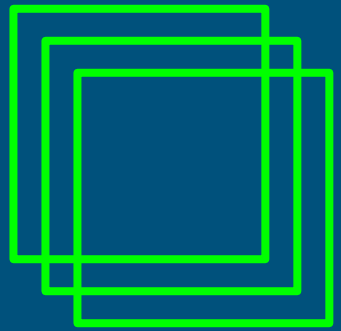
Contents

Objective

- Topic modeling on inflation tweets can help gain insights/understand people's sentiment and opinions regarding inflation.
- By analyzing the topics that comes from the tweets, businesses/entities can understand the issues that are most important to population, the concerns they have about inflation, and the ways in which they are being impacted by it.

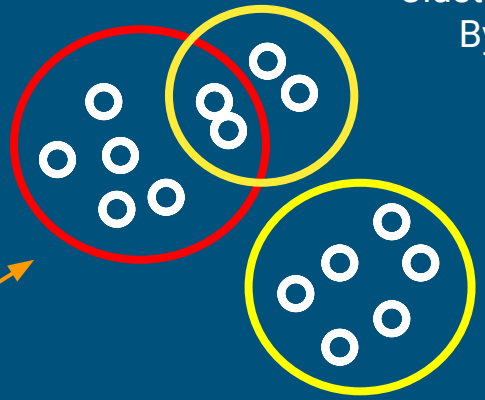


Topic Modelling (LDA)

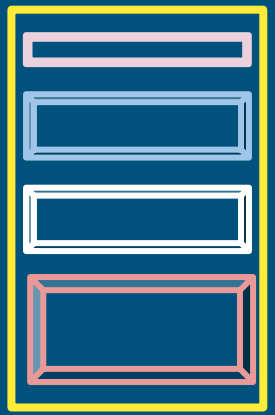


Collection Of Text Documents

Topic Modelling



Cluster Of Words By Topics



Cluster Of Document By topic

SPARK ML - Latent Dirichlet Allocation (LDA)

- Works better with large corpus
- Better results from Gensim library than spark
- Coherence score :- The quality of the topics generated by the model and to select the optimal number of topics for the model.
- Perplexity :- How well a probability distribution or model predicts a sample/ documnet

```
# Splitting the data into training and testing sets
train_data, test_data = df_clean.randomSplit([0.8, 0.2], seed=42)

# Preprocessing steps (Transformatin steps)
regexTokenizer = RegexTokenizer(inputCol="tweet", outputCol="tokens", pattern="\\s+")
stopwordsRemover = StopWordsRemover(inputCol="tokens", outputCol="filtered_tokens")
ngram = NGram(n=2, inputCol="filtered_tokens", outputCol="ngrams")
countVectorizer = CountVectorizer(inputCol="ngrams", outputCol="rawFeatures")
idf = IDF(inputCol="rawFeatures", outputCol="features")

# Preprocessing pipeline
pipeline = Pipeline(stages=[regexTokenizer, stopwordsRemover, ngram, countVectorizer, idf])

# Fitting the pipeline - training data
model = pipeline.fit(train_data)

# Transforming the training and testing data using the fitted pipeline
train_transformed = model.transform(train_data)
test_transformed = model.transform(test_data)

# LDA model (Spark inbuilt)
lda = LDA(k=20, maxIter=10, featuresCol="features")

# Fitting the LDA model to the transformed training data
lda_model = lda.fit(train_transformed)
```

Prediction/Interpretation

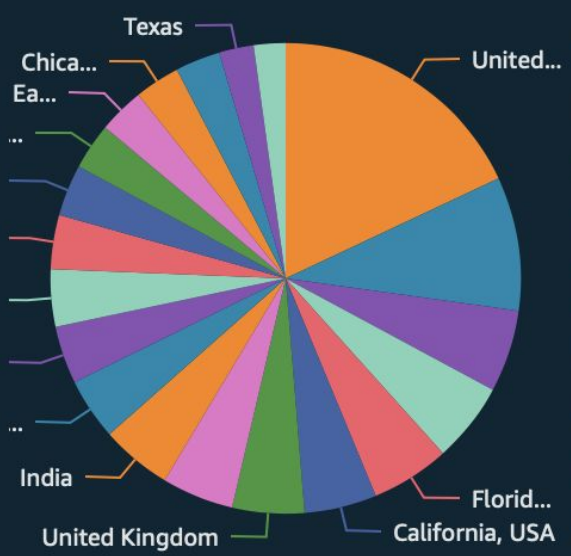
Possible interpretation from term of topics

- topic 0 - Inflation concerns India/ politics and government policies.
- topic 1 - Inflation control strategy.
- topic 2 - Corporate profitability analysis.
- topic 4 - Inflation trending downwards.

```
Topic 0: ['keeping inflation', 'money pay', 'wages keeping', 'pay nhs', 'plenty money', 'rt plenty', 'reasonable wages', 'workers r
easonable', 'nhs workers', 'president biden']

Topic 1: ['rt absolutely', 'current rate', 'outlaw cash', 'trying outlaw', 'cash transaction', 'absolutely tyrannical', 'eu tryin
g', 'rate infla', 'tyrannical eu', 'transaction current']

Topic 2: ['us inflation', 'inflation falls', 'fight inflation', 'breaking us', 'falls lower', 'lower expectations', 'rt fight', 'rt
breaking', 'inflation fathom', 'new way']

Topic 3: ['pay rise', 'nhs staff', 'rt tories', 'rt like', 'like presidents', 'one trading', 'charge one', 'presidents charge', 'ri
se pay', 'trading card']

Topic 4: ['rt inflation', 'inflation gas', 'interest rates', 'still high', 'inflation inflation', 'prices plummeting', 'republicans
talk', 'weird republicans', 'talk things', 'plummeting weird']

Topic 5: ['inflation n', 'dealing inflation', 'reminder republicans', 'dr fauci', 'busy pronouns', 'zero plans', 'still zero', 'rep
ublicans busy', 'plans dealing', 'pronouns dr']

Topic 6: ['reduce inflation', 'price increases', 'act nothing', 'annual inflation', 'inflation eases', 'inflation fallen', 'like in
flation', 'prices falling', 'food price', 'increases slowing']
```

Count of Tweets by Location

SHOWING TOP 20 IN LOCATION

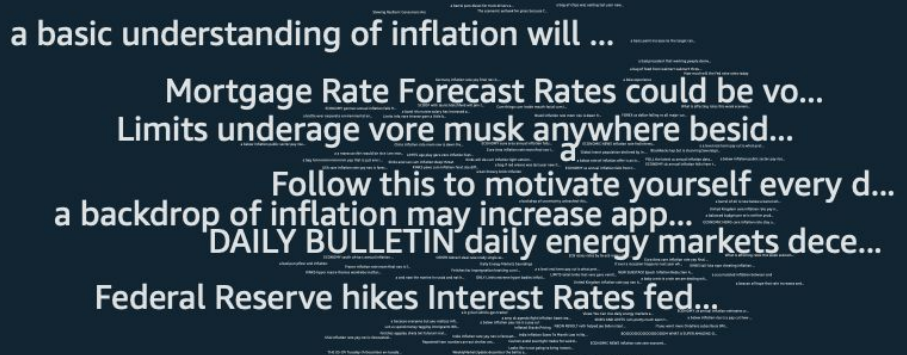


Location

- United States
- USA
- London
- London, England
- Florida, USA
- California, USA
- United Kingdom
- Canada
- India
- Texas, USA

Count of Tweets by Tweets

SHOWING TOP 100 IN TWEETS



Benefits & Drawbacks

- FAST Computing
- Scalability
- Flexibility

- Limited algorithms
- Lack of transparency
- Evaluating metrics

Challenges

- Coherence Score (More Time)
- Interpreting results
- Saving to S3 bucket
- Creating Athena queries for parquet files (udt)

Thank You