# Natural Language Processing and Unsupervised Learning Project - Song Recommendation

Calvin Yu

# Motivation


- Interested to understand the concept of Content-Based Recommender System

# Data

## Song Lyrics Dataset(Kaggle)

**data.csv**

- 170653 rows with 18 columns
- feature highlight : tempo, acousticness, danceability, key

**lyrics-data.csv**

- 379931 rows with 5 columns
- feature highlight : Lyric , SName (song name)

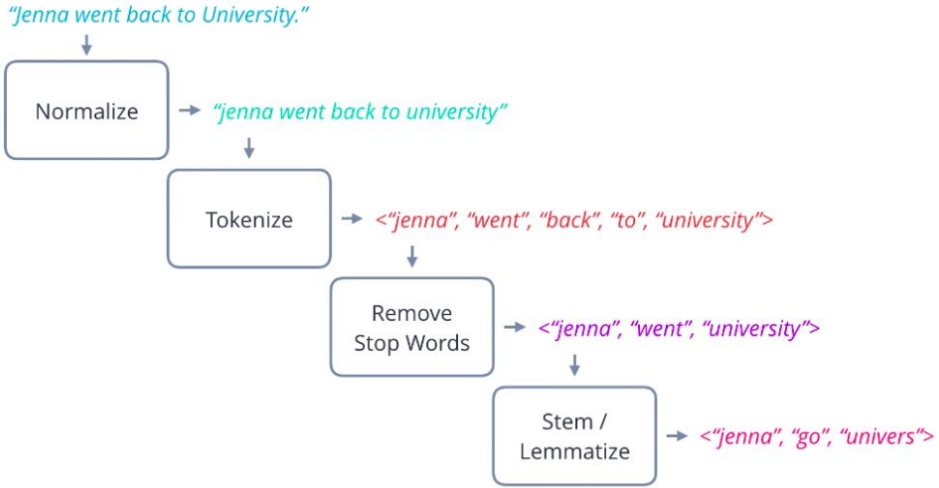
**genres_v2.csv**

- 42305 rows with 21 columns
- feature highlight : Genre, song_name

# Exploratory Data Analysis

- Brief EDA to understand the columns of each dataset
- Remove columns that is not necessarily needed
- Drop missing and duplicate values

# Word pre-processing on lyrics

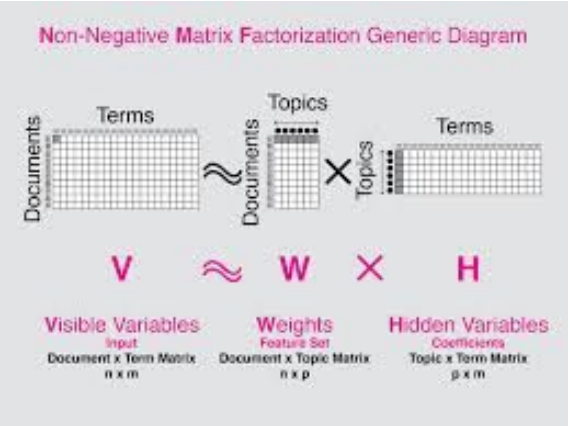



- Create a Lemmatization tokenizer
- Create a stopword list with personal extension to drop meaningless words
- Fit them to the tfidf Vectorizer (fit-transform)

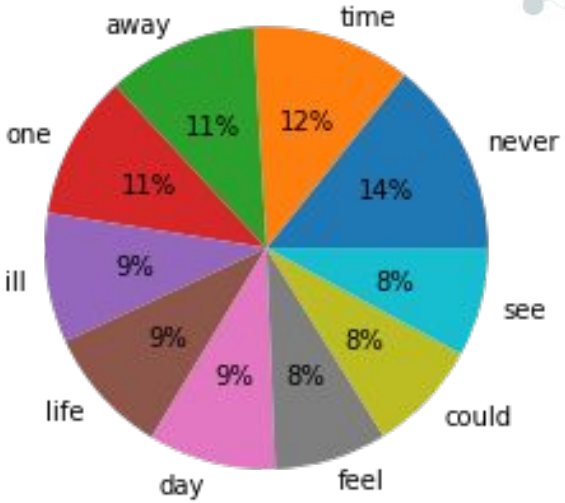
# Text Analysis

Non-Negative Matrix Factorization (NMF)

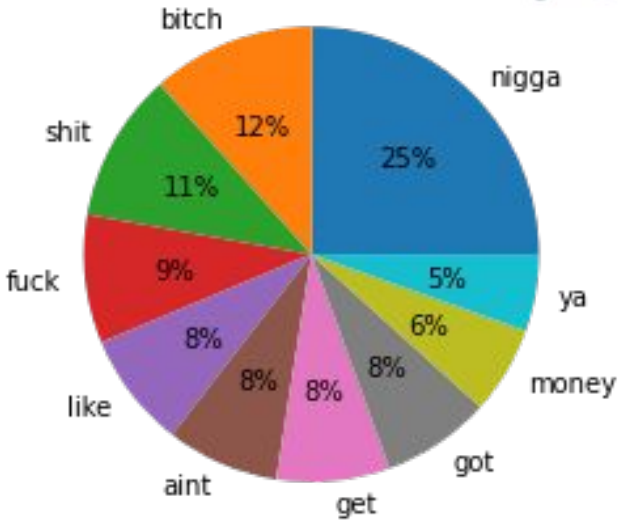
6 Components (personal pref)



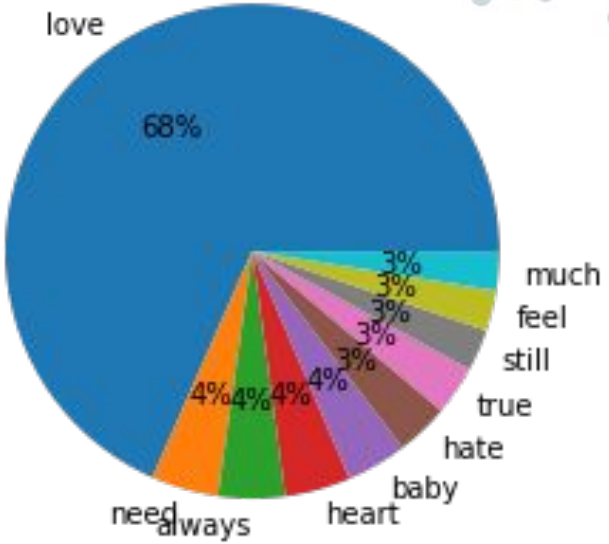
# Topic 1 - Regret



# Topic 2 - Violence

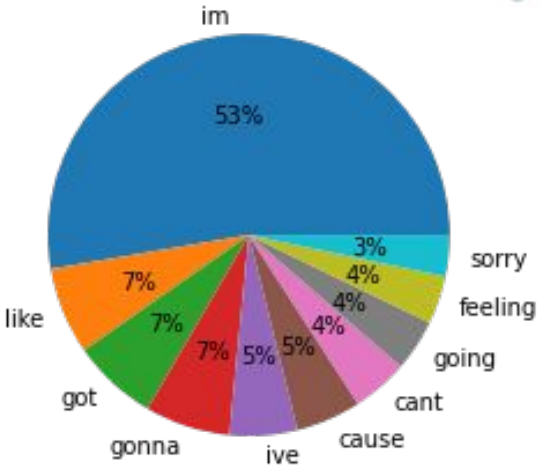


# Topic 3 - Love

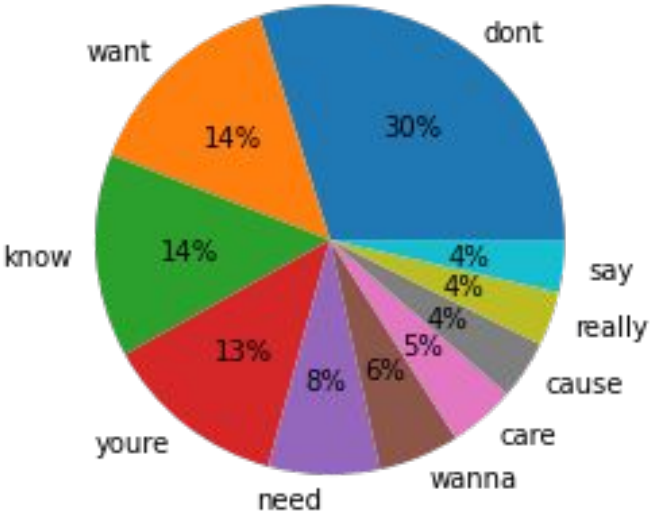

love
68%
much
feel
still
true
hate
baby
heart
always
need
3%
3%
3%
3%
3%
4%
4%
4%
4%

# Topic 4 - Relationship

# Topic 5 - Personal needs



# Topic 6 - Apology



# Song Recommendation

- Each song lyric has its unique topic score on each topic
- Combine the topic scores with the other song features, get  dummy variables for the categorical features
- Apply Cosine Similarity on the finalized dataset

# Song Recommendation

```
recommend_song("talking to the moon")
```

| | name | artists | decade |
|---|---|---|---|
| **1392** | Let Me Down Slowly | ['Alec Benjamin'] | 2010s |
| **2715** | Kids In The Dark | ['All Time Low'] | 2010s |
| **2222** | Halfway Gone | ['Lifehouse'] | 2010s |
| **1965** | Ghost Of You | ['5 Seconds of Summer'] | 2010s |
| **2688** | Ashley | ['Escape the Fate'] | 2010s |
| **1026** | Marry You | ['Bruno Mars'] | 2010s |
| **2283** | Hold Me Down | ['Halsey'] | 2010s |
| **2831** | Fallout | ['Marianas Trench'] | 2010s |
| **1847** | Warrior | ['Disturbed'] | 2010s |
| **1938** | Sidewalks | ['The Weeknd', 'Kendrick Lamar'] | 2010s |

# Further steps

- Add more songs
- Try to get user information so I can do a collaborative approach
- Try bi-grams on the lyrics