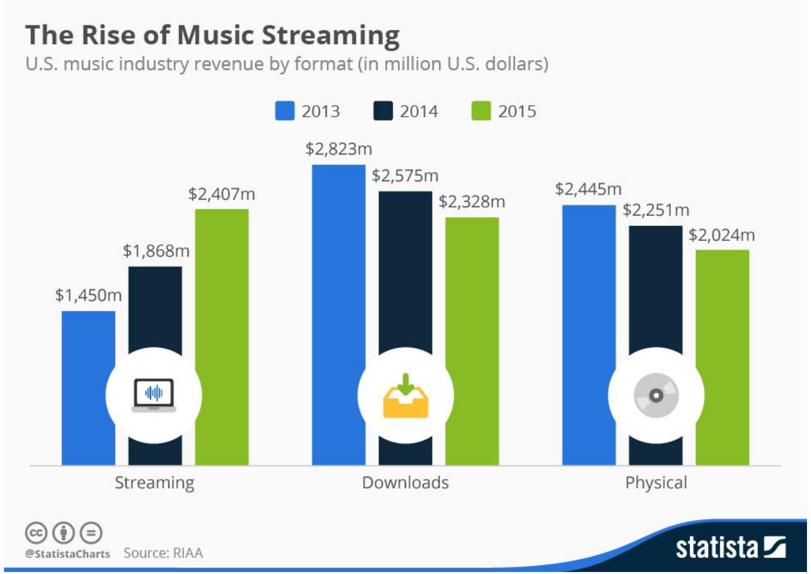
Power of Words

Lyric-based music recommendation

-- Heidi Chen, Wenxuan Dai, Xindi Zhao, Yijun Zhou

Background - Digital Music and Online Streaming

- Digital music has become the most popular source for music publication and sharing. Online streaming is among the key services provided.



The Rise of Music Streaming

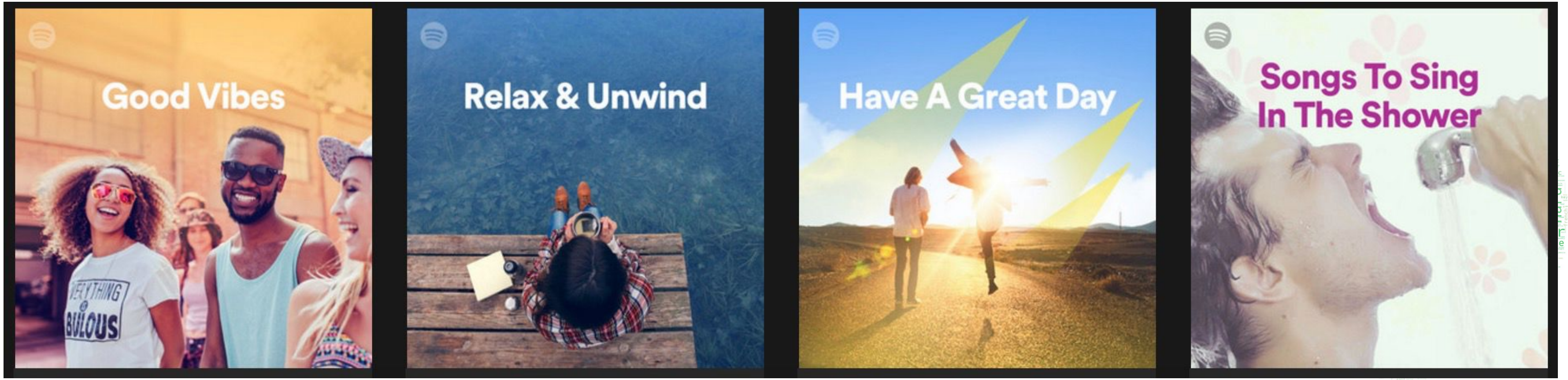
U.S. music industry revenue by format (in million U.S. dollars)

Source: RIAA

Richter, Felix. "Infographic: The Rise of Music Streaming." *Statista Infographics*. N.p., 29 Mar. 2016. Web. 21 Feb. 2017.

Background - Spotify

- Founded in 2008, Spotify provides millions of songs covering a full spectrum of music to users all over the world on multiple platforms.
- Key mission: help people find the right music at every moment through tailored song recommendations and playlists.



Motivation

“People listen to songs, or other kinds of music with text, constantly--using messages found in the lyrics to get excited, to be soothed, to express love, to help with a task, to help them cry, or to solidify the most fundamental philosophies of their lives.”

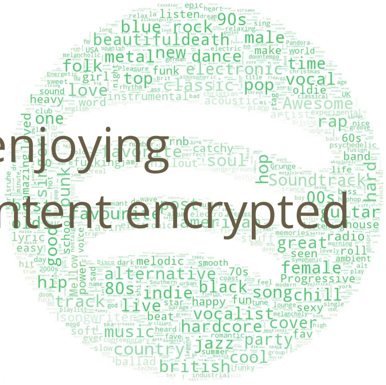
Power, Ian. *'More Than Words': Analyzing Popular Music Beyond the Lyrics* (n.d.): n. Page
http://isites.harvard.edu/fs/docs/icb.topic1089028.files/PowerSyllabusDisplay.pdf

Goal

- Analyze song lyrics to associate songs with relatable concepts such as moods, occasions, and themes.

- Create a method based on lyrics to produce playlists given different emotions and purposes, i.e. after break-up songs, relaxing music, party mix, etc.

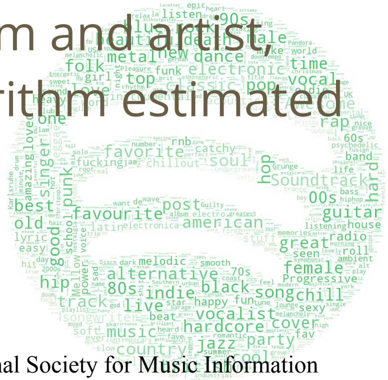
- Help create an interactive and highly personalized music enjoying experience for the users, leveraging the rich emotional content encrypted in song lyrics as well as additional song features.



Dataset Available

Million Song Dataset (MSD)

- The core of MSD is the feature analysis and metadata for one million songs.
- Derived audio features include sample rate, duration, loudness, energy, etc.
- Other metadata include information about the song, album and artist, such as releasing date, artist location. There are also algorithm estimated features: artist familiarity and artist hotness.



Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. "The Million Song Dataset". In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

Dataset Available

Last.fm Dataset - tags

- Tags generated by users from Last.fm API.
- 33,355 different tags for 9,330 songs in the subset, including information about genre, emotion, occasion, and etc.
- Examples: “rock”, “happy”, “chill”, “dance”, “00s”, etc

MusiXmatch (MXM) Dataset - lyrics

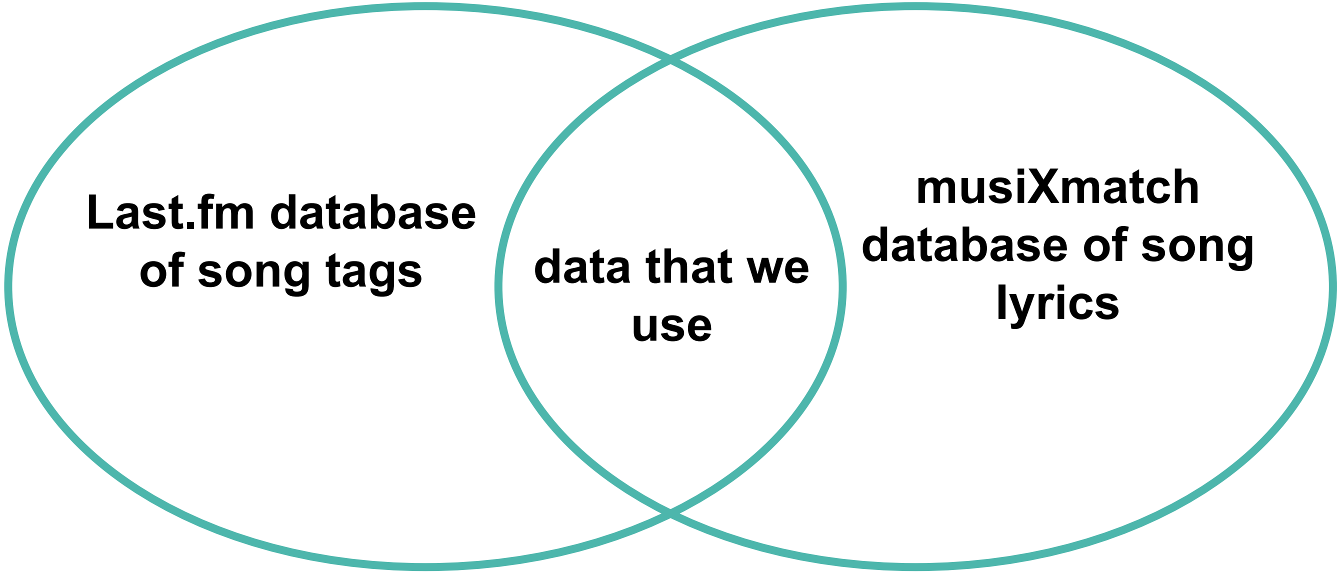
- Provides lyrics for 77% of the MSD tracks.
- Bag-of-words format: each track is described as the word-counts for a dictionary of the top 5,000 words across the set.
- Stemmed words

blues rock 90s
beautiful death male
metal new dance
folk electronic time
girl top classic pop vocal
love instrumental acoustic Awesome
heavy word oldie
one rap 60s
loved singer punk favorite catchy soul hop
fucking chillout Soundtrack hard
amazing best post 00s
favourite american guitar
old good school latin electronica house
lyric great radio
easy melodic smooth roll
alternative 70s female Progressive
hip 80s indie black song chill
track live happy fun vocalist sexy
songwriter beat hardcore cover
soft music party fav
country jazz cool

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. "The Million Song Dataset". In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

Data sets available

- Lyrics and tags can be matched using MSD track IDs.



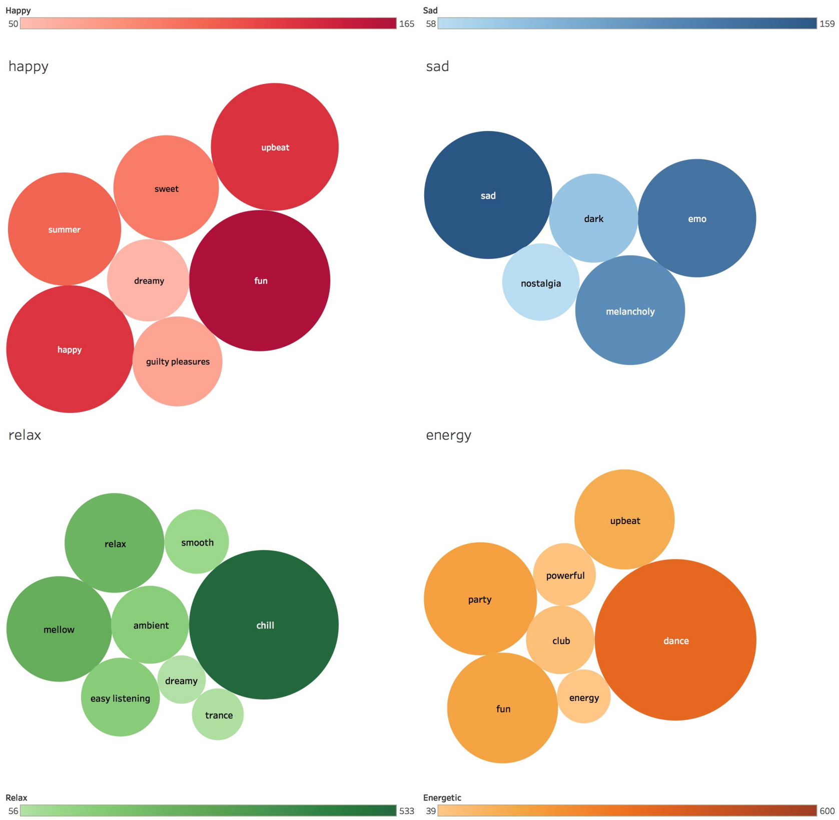
Last.fm database of song tags

data that we use

musiXmatch database of song lyrics

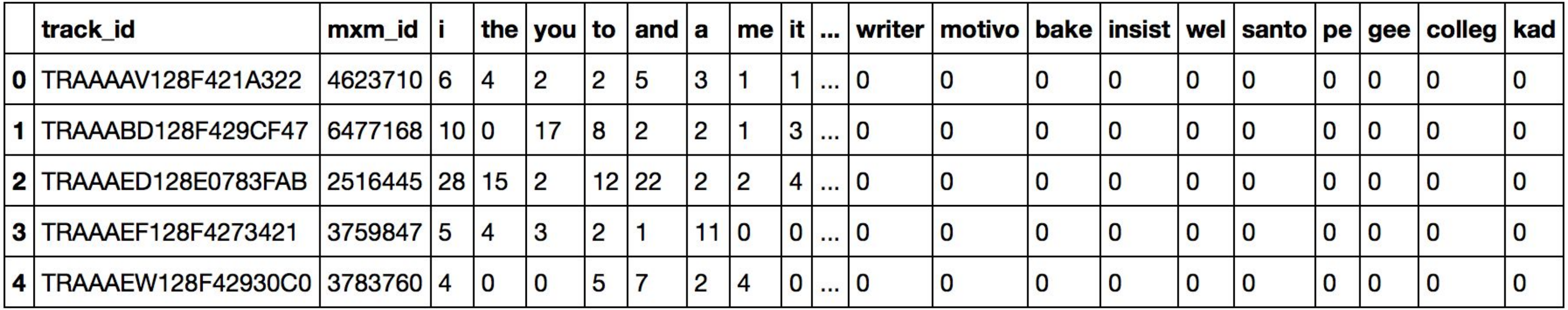
Data Exploration

- **Tags:**
  - Taking a subset of 9,330 songs with tags as an example, we see
    - More than one tag for one song, 33,355 different tags in total.
  - If we group the full dataset into 4 major mood tags (happy, sad, relaxing, and energetic) to start with we get
    - A total of 839,122 songs in the training dataset.
    - 11,123 happy songs, 11,262 sad songs, 9,198 relax songs, and 5,590 energetic songs.



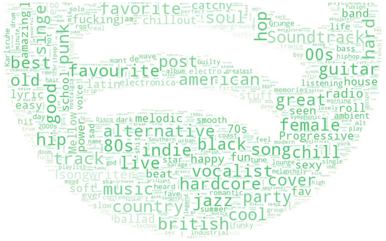
Data Exploration

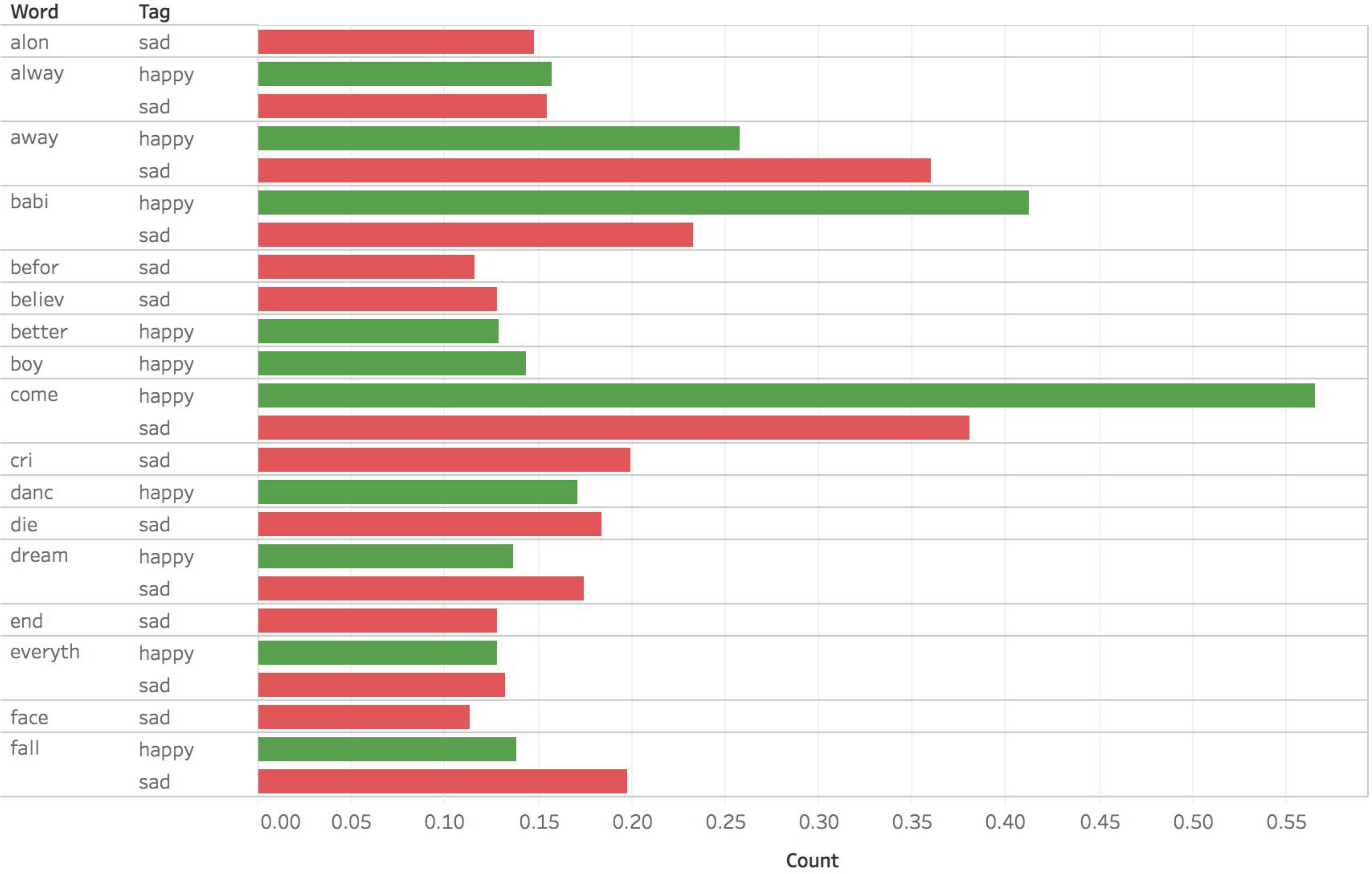
- Lyrics - Bag of words:
  - Transform the dataset to a dataframe in Python:



	track_id	mxm_id	i	the	you	to	and	a	me	it	...	writer	motivo	bake	insist	wel	santo	pe	gee	colleg	kad
0	TRAAAAV128F421A322	4623710	6	4	2	2	5	3	1	1	...	0	0	0	0	0	0	0	0	0	0
1	TRAAABD128F429CF47	6477168	10	0	17	8	2	2	1	3	...	0	0	0	0	0	0	0	0	0	0
2	TRAAAED128E0783FAB	2516445	28	15	2	12	22	2	2	4	...	0	0	0	0	0	0	0	0	0	0
3	TRAAAEF128F4273421	3759847	5	4	3	2	1	11	0	0	...	0	0	0	0	0	0	0	0	0	0
4	TRAAAEW128F42930C0	3783760	4	0	0	5	7	2	4	0	...	0	0	0	0	0	0	0	0	0	0

  - Match the lyrics with tags based on track_id:
    - Happy - 5,312, sad - 6,357, relax - 2,647, energetic - 2,647





Sum of Count for each Tag broken down by Word. Color shows details about Tag. The view is filtered on Word, which keeps 17 of 115 members.

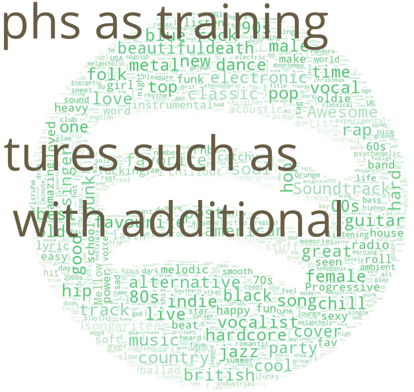
Remove stop words, sum for each tag and normalize the counts

$$Normalize_w = \frac{original\ count_w}{\max count}$$

for w excluding stopwords

Approach and Measure

- We will start with existing songs tagged with “emotion” type tags and matched with bag-of-words lyrics as our preliminary set for training models - approximately 16,000 songs for the four selected tags.
- Potentially grouping more correlated tags into the target tags will give us more data to use for emotion training.
- We will seek to incorporate more tagged articles/paragraphs as training data for language to emotion training.
- If time permits, we might incorporate additional song features such as artist and genre into the process and evaluate the model with additional criteria such as song popularity.



Relevant Knowledge: Methods I

- Support Vector Machine (SVM)

- classifies a new instance of a document D (lyrics) into a finite set C of predetermined classes (tags).

- Naive Bayes

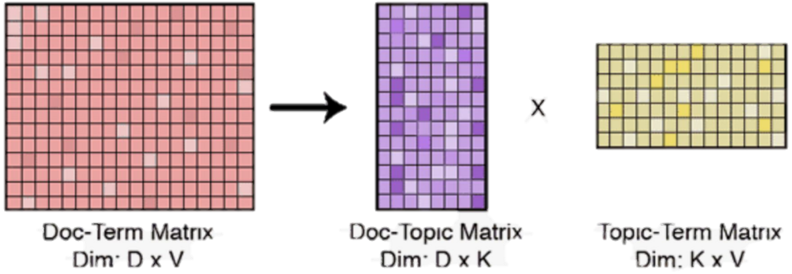
- *P(C)* is the prior probability of category C and *P (W|C)* is the conditional probability for word W given category C

$$Best = argmax_c \frac{P(W|C)P(C)}{P(W)}$$

Relevant Knowledge: Methods II

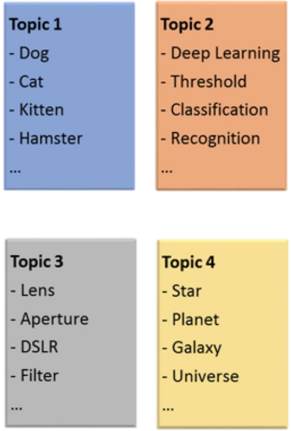
- supervised Latent Dirichlet Allocation(sLDA)

- lowering the documents’ dimensionality
- document-term vectors → documen-topic vectors
- allowing term variability represented at a topic level rather than at the raw word level.



- Probabilistic Latent Semantic Analysis (pLSA)

- the analysis of two-mode and co-occurrence data
- using words frequency as a characteristic vector
- calculating the distance in the vector space to see the semantic closeness.

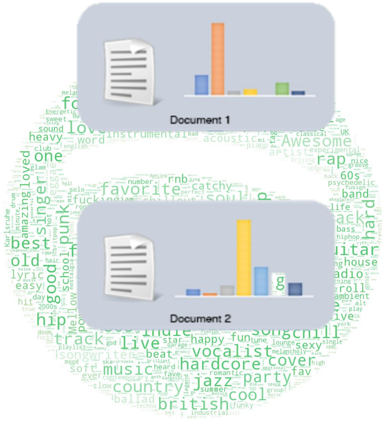


Topic 1
- Dog
- Cat
- Kitten
- Hamster
...

Topic 2
- Deep Learning
- Threshold
- Classification
- Recognition
...

Topic 3
- Lens
- Aperture
- DSLR
- Filter
...

Topic 4
- Star
- Planet
- Galaxy
- Universe
...

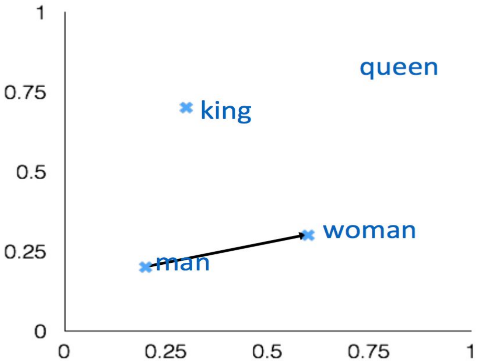


Bhadury, Arnab. "Clustering Similar Stories Using LDA — Flipboard Engineering." *Flipboard Engineering*. N.p., 8 Feb. 2017. Web. 20 Feb. 2017.

Relevant Knowledge: Methods III

Word2vec

- Word embedding
- Distributional similarity based representations
- Capturing dimensions of similarity as linear relations
- Encoding meanings in vector differences



+	king	[0.30 0.70]
–	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]

Brown, Taylor W. "Introduction to Word Embedding Models with Word2Vec." *Introduction to Word Embedding Models with Word2Vec*. N.p., 11 July 2016. Web. 20 Feb. 2017.
Manning, Christopher. “Compositional Deep Learning”, http://nlp.stanford.edu/manning/talks/NAACL2015-VSM-Compositional-Deep-Learning.pdf

Deliverable

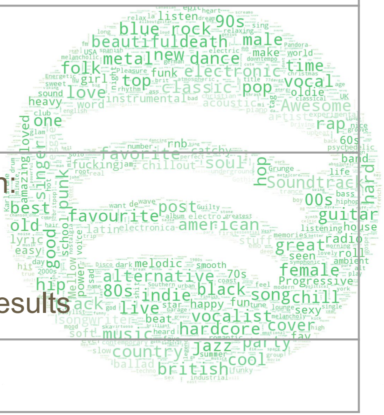
Deliverable 1	Predictive **model** trained on lyrics and existing articles which: ● Predicts theme/mood tags of an input song ● Lists a set of songs with given tags in the tag database This will be built with word2vec, NLTK(Natural Language Toolkit), LDA, etc.
Deliverable 2	Advanced predictive **model** which: ● Creates a separate model to generate the similar tags that relate to the input tag(not in the tag database) ● Based on given tag, generates a list of songs that relate to the tag according to the songs tag distribution and possibly songs' popularity
Deliverable 3	Python **module** which: ● Can be run as a standalone script and ready for demo purpose **Website** which: ● Can show the result and workflow of the entire project

Timeline - Past Milestones

Sprint ending	milestone or goal
2017-02-07	● Project set up ○ Private git repository created, TF and professor shared ○ Team communication channel (Slack) selected, TF added ○ Project management tool selected (Github), TF added
2017-02-14	● Set up Goals & Data Exploration ○ Decide on final goal of the project ○ Explore MSD, Spotify API, MXM API, Genius API ○ Confirm data on hand is adequate for the rest of the project ○ Complete first draft of scope document and send to Client for review and approval
2017-02-21	● Data scraping/cleaning & Tool/Method Learning ○ Extract data from Spotify API and MSD ○ Preliminary data visualization, compile list of technical/data and business questions for Client ○ Research into potentially relevant machine learning and NLP algorithms ● Prepare presentation and Scope of Work for Midterm 1

Timeline - Future Milestones

Sprint ending	Tentative milestone or goal
2017-03-07	● Learn selected algorithms and toolkits ○ Word2vec (Continuous Bag of Words), LDA and NLTK toolkit ○ Confirm what methods are appropriate given our existing data and goal ● Preliminary classification experiment with own data
2017-03-21	● Analyze and report model fitting progress for tags prediction ○ Further process the data into the desired form for different algorithms ○ Train and test on existing tags for model accuracy
2017-04-04	● Extend limited tags to groups of similar tags to perform lyrics-to-tag analysis ● More visualization with fitted predictive models ● Prepare for Midterm 2 presentation to class and Partner
2017-04-25	● Incorporate additional song features to further improve emotion/theme prediction: ○ Could potentially use genre, artist information for better ● Finalize models and deliverables for final presentation ● Organize codes and create demos for predictive models and recommendation results
2017-05-02	● Poster and presentation preparation and review



Citation

Bhadury, Arnab. "Clustering Similar Stories Using LDA — Flipboard Engineering." *Flipboard Engineering*. N.p., 8 Feb. 2017. Web. 20 Feb. 2017.

Brown, Taylor W. "Introduction to Word Embedding Models with Word2Vec." *Introduction to Word Embedding Models with Word2Vec*. N.p., 11 July 2016. Web. 20 Feb. 2017.

Manning, Christopher. “*Compositional Deep Learning*”, Workshop on Vector Space Modeling for NLP 2015, http://nlp.stanford.edu/manning/talks/NAACL2015-VSM-Compositional-Deep-Learning.pdf

Power, Ian. *'More Than Words': Analyzing Popular Music Beyond the Lyrics* (n.d.): n. pag. Web.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. "The Million Song Dataset". In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.