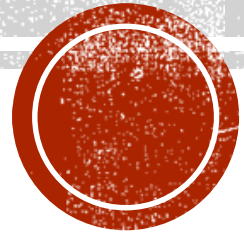
ADS PROJECT 4
WORDS 4 MUSIC



Chenxi (Celia) Huang

AGENDA

- The Project
- My Approach
- My Findings
- Concerns & Considerations

THE PROJECT

- **The Data**

The million song data
"- /metadata, -/musicbrainz, -/analysis/songs" not provided in the test data.

- **The Goal**

Based on the association patterns identified, we will create lyric words recommender algorithms for a piece of music (using its music features).

- **Evaluation Criteria (my translated version)**

  Error = mean(predicted ranks) – mean(actual ranks) of respective words

MY APPROACH

- **Part I: Feature Engineering**
  1) Feature Creation
  2) Feature Selection

- **Part II: Model Selection**
  1) Baseline
  2) Clustering (K Means and Hierarchical Clustering)
  3) Topic Modeling

PART I: FEATURE ENGINEERING

- **Feature Creation**
  - initially as many as possible
  - generate comprehensive statistics {Psych}
    e.g. "vars, n, mean, sd, median, trimmed, mad, min, max, range, skew, kurtosis, se“
  - (16-1)*13 = **195 features** in total for each song

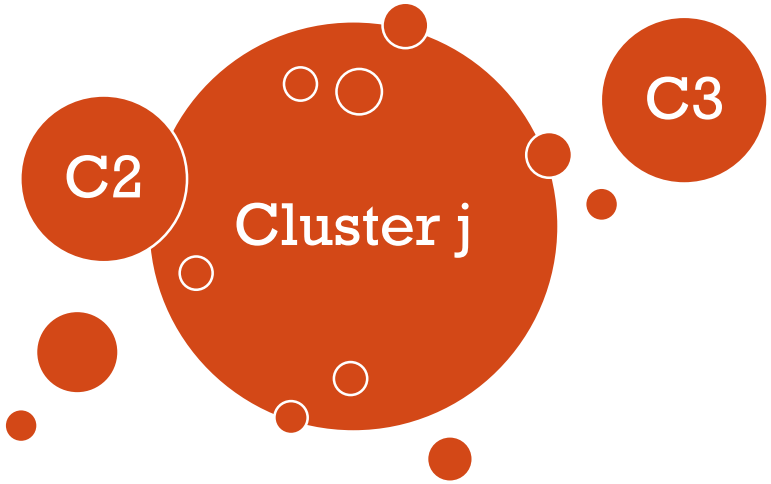
- **Feature Selection**
  - Data Cleaning (NA values, columns with mean=Inf or –Inf, etc)
  - PCA? (dimension reduction)
  - Random Forrest? (supervised learning)
  - Left with: 174 vs 153 features per song → 174 won

MODEL SELECTION

1. Baseline

   Simple: based on all songs and their word rankings in lyr.data.

2. Clustering (why?)
   - Have used it before, comfortable with it
   - Good tutorial. Clear instructions
   - Essentially the same idea as Baseline.



MY FINDINGS

1. Baseline Model

```
> # error = mean(predicted ranks) - mean(actual ranks in the test data)
> #average error is 190.3119
> mean( cv.error)
[1] 190.3119
```

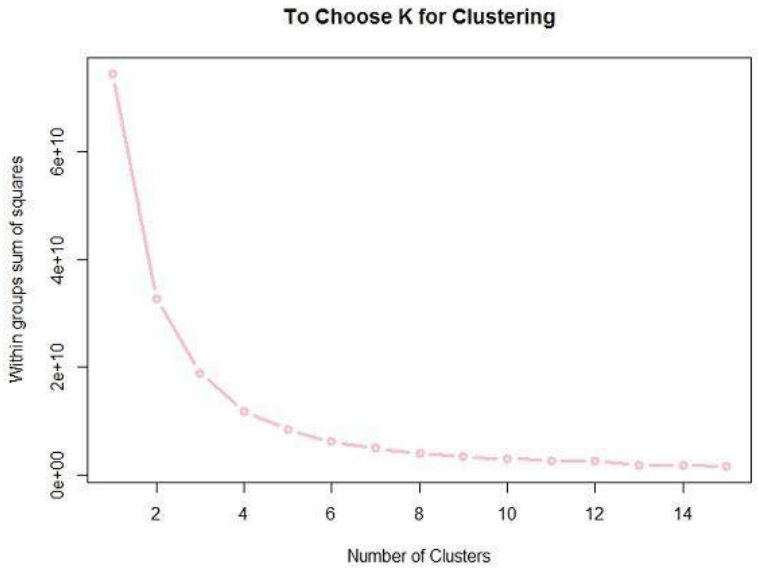
Cross Validation Results for the Baseline Model

	K=1	K=3	K=5	K=10
Average CV.Error	549.6087	549.56	549.6087	549.6087

It's not a very good result. But a pretty consistent one.

MY FINDINGS

2. K Means: How to choose K



How to choose K for K means Cluster?

Method	K	Predicted Test Topics	Interpretation
K Means	3	Topic 2 for all 100 test songs	Essentially the baseline model
K Means	5	Topic 3 for all 100 test songs	Essentially the baseline model
K Means	8	Topic 7 for all 100 test songs.	Essentially the baseline model
K Means	10	Topic 5 for 87 songs, Topic 7 for 13 songs.	Better, still not very well classified.
K Means	12	Topic 3 for 87 songs, Topic 7 for 13 songs.	Better, still not very well classified.
K Means	15	Topic 9 for 58 songs, Topic 13 for 42 songs.	Some classification, more balanced but weak.
K Means	20	Topic 5 for 42 songs and Topic 8 for 58 songs.	Some classification, more balanced but weak.
K Means	30	Topic 12 for 42 songs and Topic 22 for 58 songs.	Some classification, more balanced but weak.

At most 2 groups. Not very good results.

MY FINDINGS

2. K Means: Cross Validation

Cross-Validation Results of K Means Clustering (10 Clusters)

	K=1	K=3	K=5
i=1	2443.312	658.4373	821.5007
i=2	-	820.7046	917.1601
i=3	-	851.1187	815.6407
i=4	-	-	820.9437
i=5	-	-	860.0891
Total Mean Error	2443.312	776.7535333	847.06686

Cross-Validation Results of K Means Clustering (20 Clusters)

	K=1	K=3	K=5
i=1	673.6772	786.7195	752.5631
i=2	-	952.7943	785.4017
i=3	-	969.5194	921.3359
i=4	-	-	747.4215
i=5	-	-	774.7456
Total Mean Error	673.6772	903.0110667	796.29356

Not very satisfactory results.

Note:
K = number of folds
i = files numbers chosen for each K

MY FINDINGS

3. Hierarchical Clustering: Cross Validation

Cross-Validation Results of (Wald) Hierarchical Clustering (10 Clusters)

	K=1	K=3	K=5
i=1	609.3154	670.251	777.5202
i=2	-	782.7894	879.2189
i=3	-	807.4677	778.0394
i=4	-	-	781.7725
i=5	-	-	814.3553
Total Mean Error	609.3154	753.5027	806.18126

Still not very satisfactory results.

Note:
K = number of folds
i = files numbers chosen for each K

CONCERNS & CONSIDERATIONS

- **Feature Selection**
  - PCA ?

    (1) principal component are used as new features, instead of the original variables

    (2) only linear relationships are considered
  - 175 features. Overfitting?

- **Clustering**
  - Not well separated
  - Even if, similar bars = similar lyrics?

- **Topic Modeling**

THANK YOU!