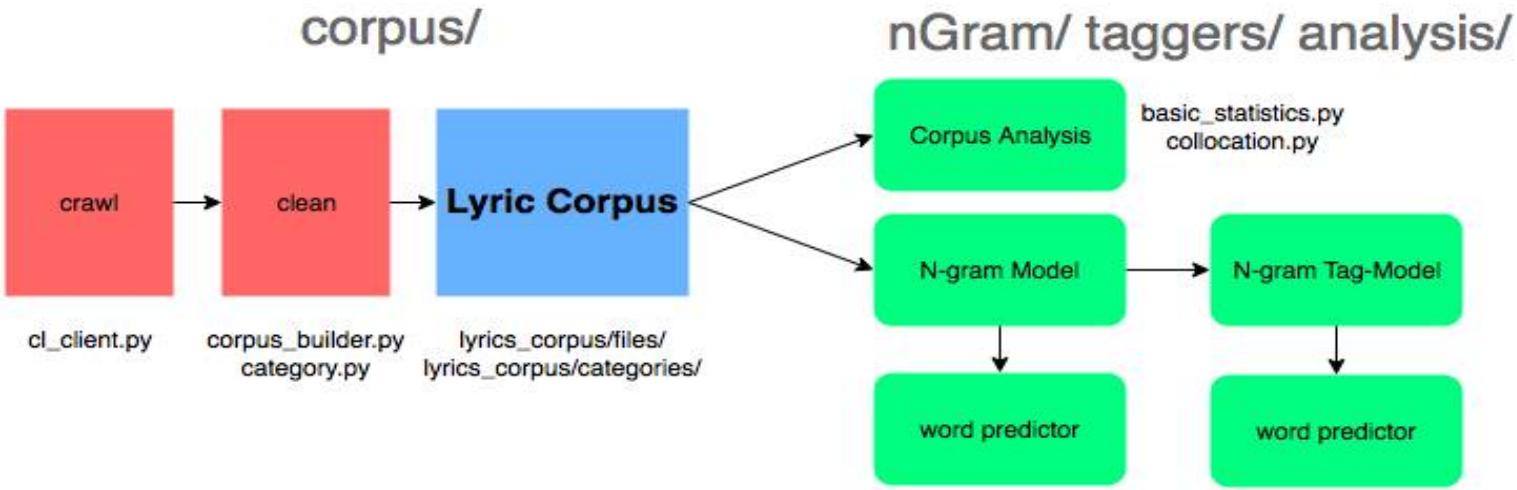
A Word Predictor For Song Lyrics

Group - 9

DD2380 Artificial Intelligence 2015 - Qi - Fu - El Alaoui - Theodoridis

System Architecture



1697 different lyrics from **59** artists

Two categories **POP** and **ROCK**

Corpus Analysis

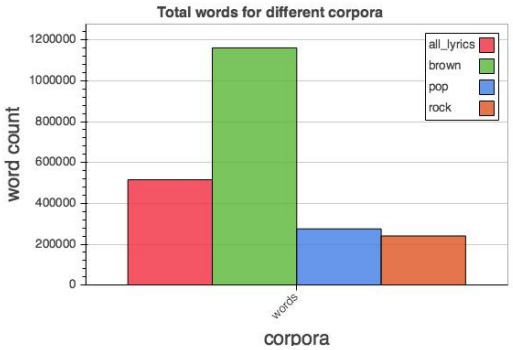
Categories are balanced in terms of

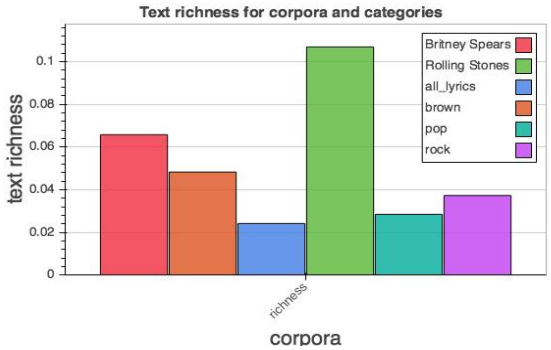
- word count
- vocabulary size
- text richness

Collocation analysis

- specific for each category
- relatively high scores

Some words and collocations only appear in a specific song. We have to consider this in training. -> shuffle





POP				ROCK		
collocation	**pmi**	**fDist**		**collocation**	**pmi**	**fDist**
kitty kat	14.26	14		arnold layne	14.41	11
raspberry beret	14.26	12		killboy powerhead	14.29	12
mamma mia	14.16	10		mi corazn	14.29	12
tick tock	13.98	14		organic anti	14.17	10
sugar coated	13.31	10		53rd 3rd	14.17	13
coucher avec moi	30.13	8		franco un american	29.62	5
christopher tracys parade	29.13	7		dub thee unforgiven	29.05	8
vous coucher avec	28.28	8		yak yak yak	28.56	5
jumping thumping shout	27.68	6		twentieth century fox	28.46	9
creole lady marmalade	27.38	5		longer shop happily	27.75	9

Table 1: Top 5 bi- and trigram collocations ranked by pmi score.

Word Predictor

- N-Gram
- Smoothing
  - Laplace
  - WittenBell
- Backoff
- N-Gram Tagging
- Linear Combination

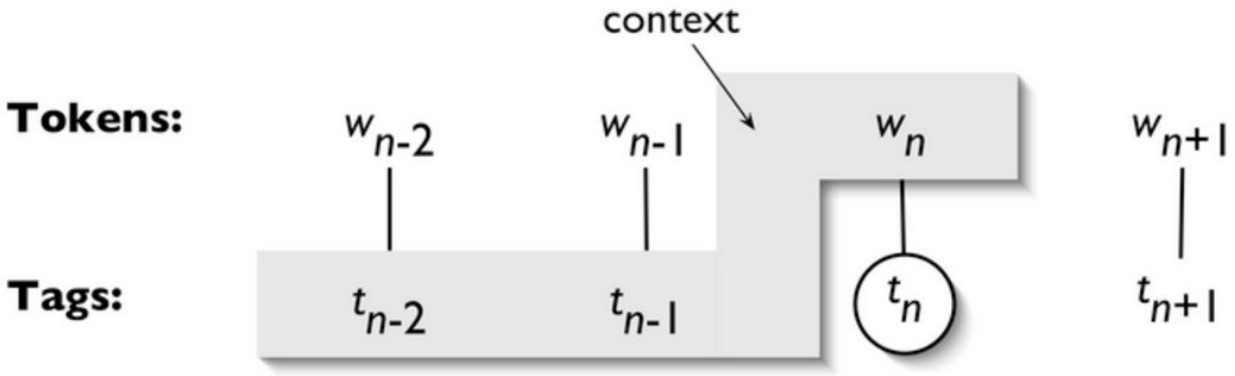
Evaluation - Random Sentence Generation

	Whole corpus	Pop corpus	Rock corpus
Bigram	leaves are the way. i am a little girl.	avant que tu amor. i am not you.	nothing to the sun. i am a little girl.
Trigram	sweet little woman get high. i am a a diva hey.	so who are you ok.	well thats alright, its just a shot away
Quadrigram	my memory reces at speeds. come on, come on.	never too much. you will be mine and i will be your light.	you made your first kill now. find another place to feed your greed.

Evaluation - Perplexity

	N-Gram with backoff	N-Gram with backoff and WittenBell discounting
Bigram	82.36	67.03
Trigram	39.08	28.75
Quardrigram	29.05	20.62

N-Gram Tagging



source: http://www.nltk.org/book_1ed/ch05.html

- Train a tagger with Brown corpus
- Tag our corpus with the trained tagger

Linear Combination

$$P(w_i|w_{i-1},t_{i-1},t_{i-2}) = \alpha \times P(w_i|w_{i-1}) + (1-\alpha) \times \max_{t_i \in T(w_i)} [P(w_i|t_i) \times P(t_i|t_{i-1},t_{i-2})]$$

- Accuracy
  - Separate corpus into training corpus(80%), test corpus(10%) and validation corpus(10%).
  - Choose 100 sentences ramdomly from validation/test corpus.
  - For each sentence, choose a position randomly and predict the word at that position.
- Choose the best N-Gram model
  - N-Gram with WittenBell discounting and backoff
- Choose the best N for words and tags
  - Given alpha = 0.5 : N for tags is 2 and N for words is 4

Linear Combination - Choosing Alpha

alpha	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Accuracy(%) for Whole Corpus	42.5	41.5	43.5	**47.5**	44.5	37	41.5	45	41.5	38.5
Accuracy(%) for Pop Corpus	45	48	44	46.5	**49**	46.5	47	40.5	39.5	44
Accuracy(%) for Rock Corpus	49.5	47	44.5	45.5	48	40	43	35.5	**51.5**	40

Accuracy(%) on the test corpus

	1	2	3	4	5	Avg
N-Gram with backoff	45	45	40	44	40	42.8
N-Gram with WittenBell discounting and backoff	43	50	48	45	46	46.4

Examples - Whole Corpus

Sentences	Context	Predicted Word
telling mother nature bout **you** and me.	[('mother', ('nature'), ('bout', 'NN')]	you
walk away and **taste** the pain.	[('walk'), ('away'), ('and', 'CC')]	taste
i can not get him off my **mind** and it scares me.	[('him'), ('off'), ('my', 'PP$')]	mind
let us keep on rocking**.**	[('keep'), ('on'), ('rocking', 'NN')]	.
you do not **need** to cause a fuss.	[('you'), ('do'), ('not', '*')]	have
every time that i look in the **mirror**.	[('look'), ('in'), ('the', 'AT')]	eye

Examples - Pop Corpus

Sentences	Context	Predicted Word
so its not just **gonna** happen like that.	[('its'), ('not'), ('just', 'RB')]	gonna
and good old boys **were** drinking whiskey and rye.	[('good'), ('old'), ('boys', 'NNS')]	were
its beyond me, i cannot **carry** the weight of a heavy world.	[(','), ('i'), ('cannot', 'MD*')]	carry
i could have another you by **tomorrow**.	[('another'), ('you'), ('by', 'IN')]	tomorrow
i do not **care**, as long as you love me, baby.	[('i'), ('do'), ('not', '*')]	know
i saw you **standing** at the gates.	[('i'), ('saw'), ('you', 'PPO')]	.