Next step: (consider normalization by the length of the lyrics when possible)

1. N gram

Tf-idf, Ngram range = (1,3) max\_features = 300

1. \* type-token ratio = # of unique words/tokens (feature type: vocabulary)
2. feature type: style:
3. POS (map to supertag V N ADV)
4. Length per line
5. \*Repetition:

letter level: # of repeated letter >3

word level: count 1 word and 2 words repetition as one feature

1. Feature type: SEMANTICS \*

Regressive Imagery Dictionary：RID classifies words as belonging to the separate fields “conceptual thought” (abstract, logical, reality-oriented), “primordial thought” (associative, concrete, fantasy), and “emotion”.

<https://github.com/jefftriplett/rid.py>

1. Feature type: ORIENTATION
2. Past tense ratio: the fraction of past tense verb forms to all verb forms as a feature
3. Self-reference ratio: # of first person singular and plural/ 2 and 3 person
4. \* the ratio of first-person singular pronouns to second person
5. \* Feature type: song structure

Chorus: To be able to recognize such cases, we compute the overall similarity between two lines as a weighted sum of their lexical and structural similarities which are modelled in terms of word and POS tag bigram overlaps, respectively. (not very clear how to implement)