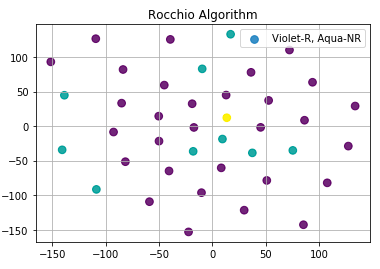
Assignment 3

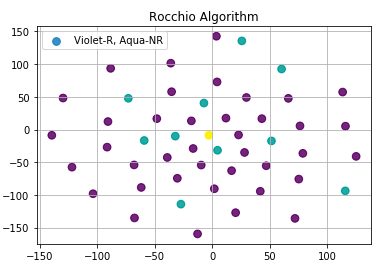
* William Scott (MT18026)

**Question 1:**

Inferences:

* After certain number of steps, the query tends to move closer to the relevant documents and farther from the non-relevant documents.
* Cosine similarity works better with TF-IDF vectors





**Preprocessing Used:**

* Lowercase
* Remove Punctuation
* Lemmatize
* Convert numbers
* Again remove punctuation
* Lemmatize
* Remove stop words
* Stem

Alpha – 0.5

Beta – 0.3

Gamma – 0.2

**Assumption:** The documents that are not marked as non-relevant are considered as relevant from the displayed top 10.

**Process:**

* Find Relevant and Non-Relevant Documents
* Compute Centroids
* Apply formula
* The query vector will tend to move closer to the relevant documents and away from the non-relevant documents.

Q\_m = alpha\*Q + beta\*Q\_R – gamma\*Q\_nr

**Statistics:**

Vocab size: ~20k

Number of Docs: 2000

To create corpus: 150 sec

To Vectorize: 30sec

To calculate cosines: 2sec

**Question 2:**

Dataset:

* Microsoft URL Dataset, in which each query-url pair is in list form.
* Each pair has information regarding the URL, and there is also relevance mentioned.
* Each pair is of 136 dimensions.

Procedure:

* Load file
* Read file
* Iterate the file
  + Split the string using space
  + Check if the first name is qid:4
  + If it is
    - Extract the 1st and 75th feature
    - 1st is the relevance, and 75th is Rank
  + Else
    - Quit
  + Store the values
* Now sort the stored valued according to the rank.
* Calculate precision and recall at every point
* Plot precision – Recall curve.

Analysis:

* Total number of relevant docs – 43
* Total qid:4 available are 103
* Recall will always keep increasing
* Precision might vary.

