CSE472: Natural Language Processing

Assignment II & III: POS Tagging

Rajesh Kumar (201307556)

Kiran Raj S R (201307593)

Abhinav Yadav (201307635)

October 22, 2013

This project consists of 2 parts:

1. POS Tagging using supervised learning.
2. POS Tagging using unsupervised learning.

We have selected ***Hindi, Telugu, Kannada & Tamil*** languages for our project.

1 POS Tagging using Supervised Learning

1.1 Assumptions

(i) Tag Set

We have used LTRC tag set which consists of 26 tags.

(ii) Start and End Markers

Each sentence has been enclosed within <s> START and </s> END markers. The role of these tags is just to avoid inclusion of invalid transitions between the sentences. And hence they don't appear in the output file.

Example:

<s> START आपके PRP हिन्दी NN पसन्द NN करने VM पर PSP खुशी NN

हुई VM </s> END

<s> START **साथ NS** Tही RP मैने PRP लेख NN पर PSP कुछ QF

बदलाव NN किए VM है VAUX </s> END

Without START and END tags, VM to NS transition (shown above in bold) could have been a valid transition, which has been ignored now.

(iii) Handling unknown words

Whenever an unknown word is encountered in test data, its emission value is 0. Due to which 0 gets propagated while creating probability matrix in Viterbi algorithm. We tried stemming for this earlier, but they don't seem to give an appreciable result w.r.t unknown words. So to avoid this problem, we have used ***emission value=0.001*** for all the unknown words encountered in testing sentences.

1.2 Task Descriptions

(i) Manual tagging

We performed manual tagging on the training corpus provided to us to get the POS tags corresponding to each word. This tagging was performed with the help of shallow parser provided by LTRC.
We cleansed the output of shallow parser to obtain the tags (by simple script) and manually cross verified the tags assigned by shallow parser.

(ii) Creating Emission Matrix

Using the data provided by the files created in step-1, we can create emission matrix as follows.

$$P\ (w_i | g_i) = \text{count}\ (w_i\ ,\ g_i)\ /\ \text{count}\ (g_i)$$

(iii) Creating Transmission Matrix

Using the data provided by the files created in step-1, we can create transmission matrix as follows.

$$P\ (g_i\ |\ g_{i-1}) = \text{count}\ (g_i\ ,\ g_{i-1})\ /\ \text{count}\ (g_{i-1})$$

(iv) Creating Viterbi Matrix

We create a 2 dimensional matrix of **k*n** size.
k: Number of tags.
n: Number of words in the test sentence under consideration.

In this matrix, rows correspond to the tags and the columns correspond to the words in the test sentence under consideration.

We use the following method to fill the values in this matrix.

$$P(g_i/w_i) = P(w_i/g_i = t^1) * \mathbf{max}(\ P(g_i = t^1/g_{i-1} = t^k) * P(w_{i-1}\ /\ g_{i-1} = t^k)\ )$$
$$\text{Where } k = 1, 2, 3 \ldots \text{ number of tags}$$

(v) Traversing the Viterbi Matrix

We traverse the matrix generated in the previous step to get the tags corresponding to each word for the test sentence under consideration.

1.3 Results

After applying the steps mentioned above, we obtained the pos tags corresponding to each word for all the test sentences (for 4 languages).

The corresponding output files can be found at,
./saved outputs/Supervised Tagging/<language>_ tags.txt

Sample output is as follows.

Hindi: हिंदी_NN के_PSP पहले_NST प्रगतिशील_JJ लेखक_NN थे_VM प्रेमचंद_NN ._.

Telugu: హిందూ_NNP సంఘం_NN ఒక_JJ కులాల_NN కూటమి_NN ._.

Kannada: ಭಕ್ತ_NN ಕನಕದಾಸ_NNP ಚಿತ್ರದಲ್ಲಿ_NN ಸಾಹಿತ್ಯದಲ್ಲಿ_NN ಸಹಾಯಕರಾಗಿ_NN ದುಡಿದರು_NN ._.

Tamil: தலித்_NN இலக்கியம்_NN பற்றி_RB பல_QF கட்டுரை_NN எழுதியுள்ளார்_VM ._.

1.4 Observations

(i) Execution time

Table: Training time for different number of training sentences.

Number of Sentences	Time taken for training (in Seconds)
200	0.1170
400	0.2899
600	0.4769
800	0.7309
1000	0.9470

Figure: Training time for different number of training sentences.

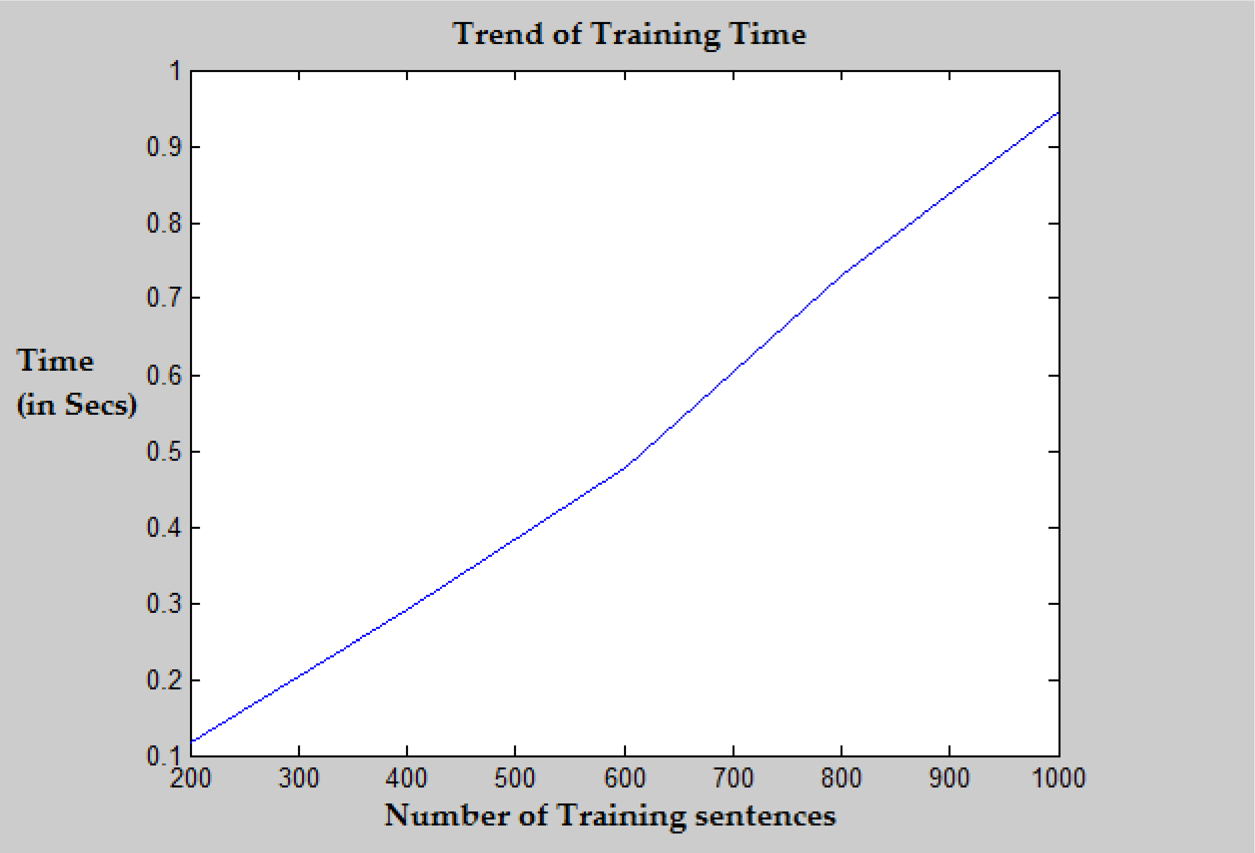
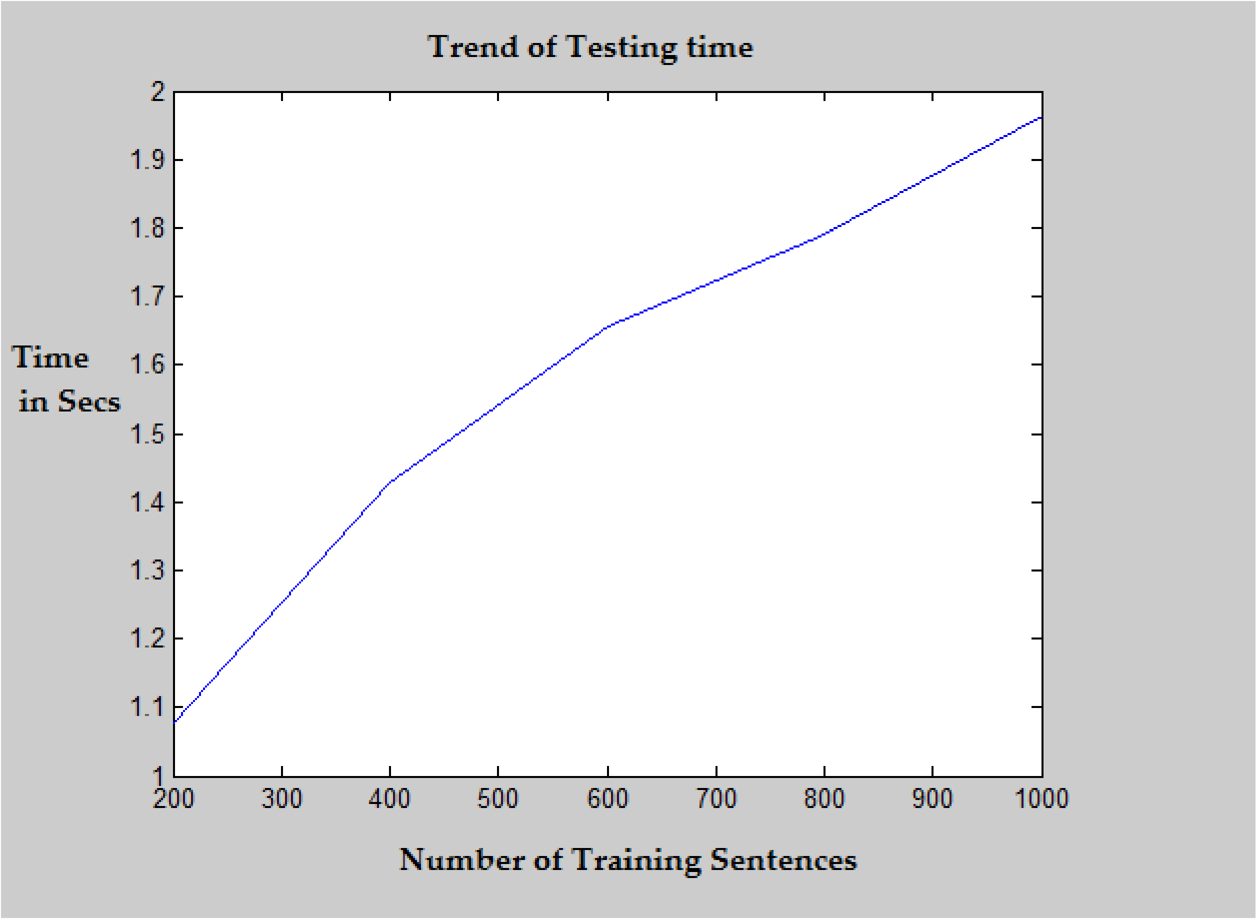


Table: Testing time for different number of training sentences.

Number of Training Sentences	Time taken for testing 100 Sentences (in Seconds)
200	1.077
400	1.4270
600	1.6559
800	1.7909
1000	1.9649

Figure: Testing time for different number of training sentences.



Average time taken to tag a single test sentence is 21.19 milliseconds.

(ii) Stemming

To handle unknown words, we tried to stem a word to make it a length n word [n=4, 6]. But number of unknown words didn't drop significantly for the given corpus. Rather it introduced a few problems as explained below.

Example: For n=4, "Import" and "Important" will give us the same stem "Impo", which is undesirable since these words have completely different meaning.

So a blunt stemming is not effective. A sensitive stemming (with linguistic rules) might have given a better result. But keeping in mind the complex linguistic properties of words in unknown languages, we preferred **NOT** to use Stemming for our project.

2 POS Tagging using Unsupervised Learning

2.1 Assumptions

(i) Number of Clusters

Number of clusters considered by us in this case =26 i.e. Number of tags in the LTRC tag set. [i.e. 1 cluster corresponding to each tag]

(ii) Context Depth

For checking the context, we have considered 4 words to the left of the given word (left context) and 4 words to the right of the given word (right context).

(iii) Initialization of vectors

To introduce randomness in the initialization of 26 centroids (while using K-means clustering), we have initialized the centroids with feature vectors of random words (picked from our training corpus).

(iv) Number of feature words

We have used top 125 words as feature words. But we have removed top 25 words (Stop Words) among these 125 words. So only the next 100 words have been considered as feature words.

(v) Distance measure

We have used euclidian distance for calculating the distance between two feature vectors.

(vi) K-Means Clustering methodology used

2.2 Task Description

(i) Feature word selection

From the given training corpus, we have considered top 100 words as feature words (after chopping off top 25 words (stop words) from top 125 words).

(ii) Creation of feature vectors

We have generated a *m x 2n* co-occurrence matrix where [m: number of words, n: number of feature words]

a. $C_{i,j}$: Count of number of times i^{th} token type occurs with the j^{th} token type as left hand neighborhood of length 4. (left context).

b. $C_{i,j+n}$: Count of number of times the same i^{th} feature token occurs with the same j^{th} token type as right hand neighborhood of length 4 (right context).

(iii) Comparing words

Each word will be represented by a feature vector. Distance between each word is given by Euclidian distance.

(iv) K-Means Clustering on feature vectors obtained.

a. Centroid initialization
b. Shuffle the centriods on the basis of average distance. Check the Error.
c. Continue till the error becomes very less.

(v) Using HMM to get the tags

Each word will belong to a cluster. Now we change the tag set used in phase-1 (supervised tagging) of our project, we can get the corresponding tags for testing data. [By application of same HMM model used in phase-1]

2.3 Results

Words in same box belong to same cluster.

We are enumerating top 10 words from any 10 non-empty clusters. [Less than 10 if the cluster does not have enough words].

Hindi:

ऋषि
प्रधान
प्रधानमंत्री
सेनानी
वकील
पत्र
कविता
कहानी
नर
सखा

भाषाओं
तमिल
शिक्षा
लेखक
उर्दू
अरबी
लिखा
ज्ञान
लेखो
अंग्रेजी

कवि
भाषा
संस्कृति
प्रथम
सम्पादन
नाटक
पत्रकार
वाणी
कोश
श्रेष्ठ

पूर्णिमा
सदस्य
अनाम
अतुल
पृष्ठ
विजय
ठाकुर
अरविन्दन्
युकेश
वर्मन
भटनागर

संस्कृत
उपन्यास
जैसे
प्रयाग
माया
विचार
काऊ
विद्वान
लगता
मूल

राज्य
उत्तर
भारतीय
सिक्किम
जिला
प्रदेश

उनके
हेतु
यात्रा
गाइड
संख्या
दो
इतिहास
करीब
त्याग

विश्वविद्यालय
ऋग्वेद
आज
मानते
छोटी
इकाई
रचना
बड़ा
जनसंख्या
पुराने

कि
प्रकार
हो
कर
तरी
मी
मे
ये
कला
देखें

सनातन
वैदिक
प्रारंभिक
शैली
संसार
अर्थशास्त्र
परमात्मा
मिलन
भरत
पंकज

Telugu:

రవాణా జాబితా పేజీకి దీని అప్లోడు విజయవాడ చాలా పాత అని ఆంధ్రప్రదేశ్	ఆ తరువాత హైదరాబాదుకు బదిలీ అయ్యాడు గ్రామంలో ప్రజల ప్రధాన వృత్తి వ్యవసాయం	నుంచి వరకు రైలు నుండి బస్సు గ్రామానికి తరచు నుంచి వరకు రైలు	కడప చిత్రం ఆస్ట్రేలియా పక్షి ఉన్నవి సమాచారం టెస్ట్ కడప చిత్రం ఆస్ట్రేలియా పక్షి
జిల్లా రాజ్ మంత్రిత్వ శాఖ వెబ్ సైటులో తాలూకాల జిల్లా రాజ్ మంత్రిత్వ శాఖ	కేవలం ఇది నాకు ఎందుకంటే ఇంజనీరింగ్ ప్రయోగాలు వికీలో అర్థం మౌలిక ఉంటుంది	వికీపీడియా వ్యాసం పేజీలో సినిమా వారపు	ఊరు వికీపీడియాలో లేదు మధ్య పేజీ మంచినీటి సంవత్సరం ఉన్నది ఊరిలో ప్రతి
ఉదాహరణకు వికీపీడియా:తొలగింపు వ్యాసాలు చర్‌చ పంట వరలక్ష్మి వ్రతం అన్న విధమైన భౌతిక	దేశం ప్రభుత్వం రాష్ట్రంలోని దేశము		

Kannada:

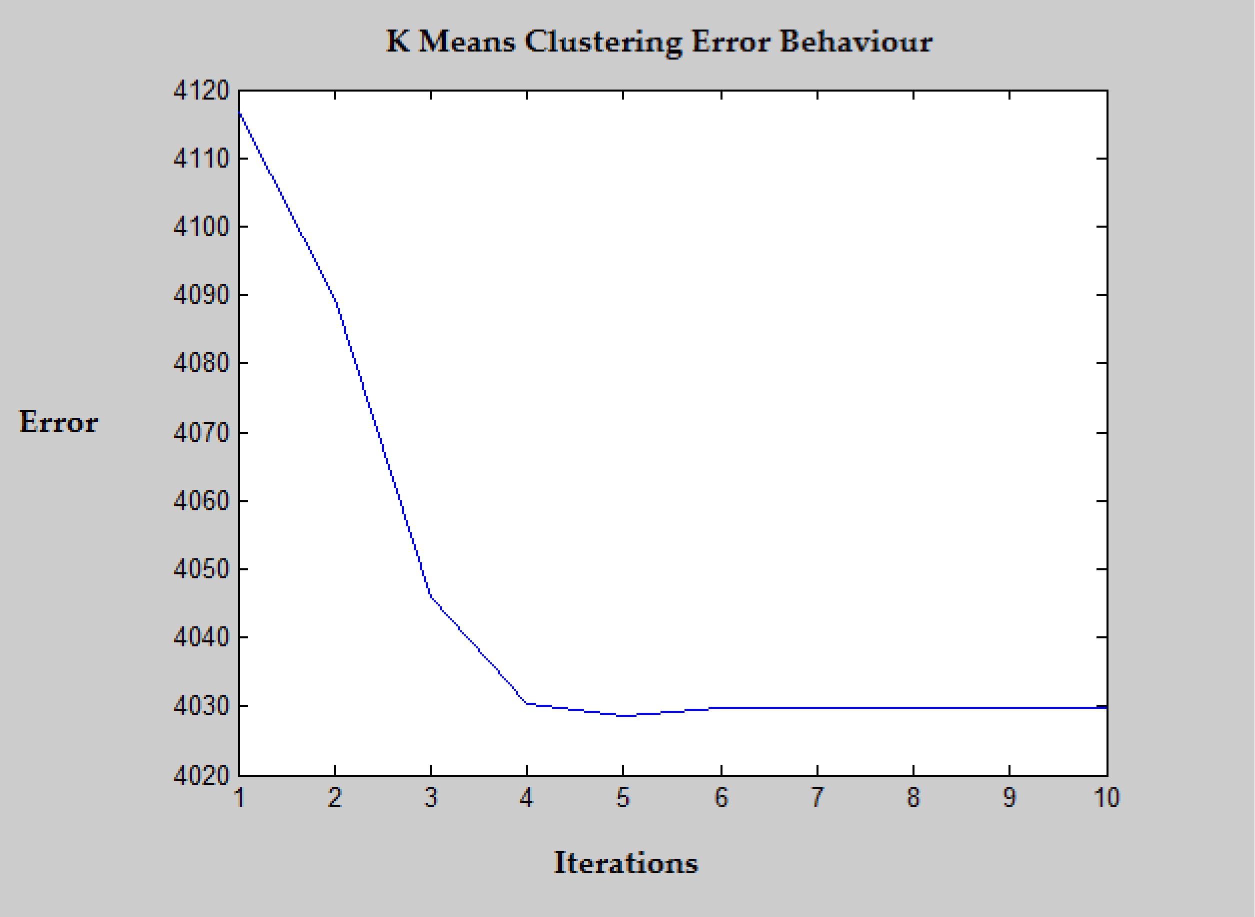
ಸಂಯುಕ್ತ ಸಂಸ್ಥಾನ ಸಂವಿಧಾನ ದಿನಾಚರಣೆ ಖಂಡದ ಅತ್ಯಂತ ಚಿಕ್ಕ ದೇಶ ಆಗ್ನೇಯ ಅನನ್ಯ	ಗಳ ಸಾಂಸ್ಕೃತಿಕ ಮೈಸೂರು ಭಾರತದ ಅನೇಕ ಆದ್ದರಿಂದ ಹಾಗೂ ಜಿಲ್ಲಾ ಕೇಂದ್ರ ಪಶ್ಚಿಮ	ರಾಷ್ಟ್ರ ವಿಶ್ವವಿದ್ಯಾಲಯವು ಇದೆ ಜಾನಪದ ತಜ್ಞ ಅಕಾಡೆಮಿ ಪ್ರಶಸ್ತಿ ರಾಜ್ಯೋತ್ಸವ ಮಹಿಳಾ ಸಂಗೀತ	ಕೆರಿಬ್ಬಿಯನ್ ಸಮುದ್ರ ಭಾಗದಲ್ಲಿ ಇರುವ ದ್ವೀಪ ಜನಸಂಖ್ಯೆ ಹೊಂದಿರುವ ಓರೆ ಅಕ್ಷರಗಳು ಹೀಗೆ ಕೆರಿಬ್ಬಿಯನ್
ಪೂರ್ವ ಕಡಿಮೆ ಇದು ಕೊಡಲಾಗಿದೆ: ಮಠಗಳು ಬಹಳ ಆದರೆ ಅವರು ತಮ್ಮ ಮಾಡಿ	ಅಮೇರಿಕ ಉತ್ತರ ಮಧ್ಯ ಚಿತ್ರ ದಕ್ಷಿಣ	ಅವರ ಹೆಸರಿನ ಇದೇ ಕಾದಂಬರಿ ಆಧಾರಿತ	ಟೆಂಪ್ಲೇಟ್ ಕೆಲವು ಮೇಲೆ ಹಲವು ಈಗ ಮಾಹಿತಿ ಬೆಂಗಳೂರು ಹುಬ್ಬಳ್ಳಿ ಪತ್ರಿಕೆಗಳಲ್ಲಿ ವಿಶ್ವಕೋಶದ
ಸಂಸ್ಥೆಯ ಸೋದರ ತಾಲೂಕು ದಲ್ಲಿನ ಧಾರ್ಮಿಕ ಗ್ರಂಥ ನಿರ್ದೇಶನದ ವರ್ಗ ಹೊಯ್ಸಳ ಒಬ್ಬ	ಫಿಲಿಪ್ಪೀನ್ಸ್ ಪ್ರಮುಖ ತಾಲ್ಲೂಕು ಶಿವಮೊಗ್ಗ ಆಫ್ರಿಕಾ ಹೊಂದಿದೆ ಪ್ರಕಾರ ಗೋಕಾಕ ದಿನ ಭಾಷೆ		

Tamil:

நகரில் தஞ்சாவூர் தொழில்கள் மலேசியா சீன இந்தியாவின் இதை ஈரோடு தொடர்பான சிவாலயமாகும்	சிந்து செல்வா விக்கிபீடியா அல்ல கோபி பயனர்கள் பாலாஜி உடைய ஆக்கத்தை அடிப்படையாகக்	பெற்ற சம்பந்தர் அப்பர் சுந்தரர் மூவரதும் பாடல்	வீச்சு என்பது மிக தமிழ்ச் சிங்கப்பூர் பண்பாடு நாட்டார் மாற்றி விடுங்கள் கலைக்களஞ்சியம்
நினைக்கிறேன் மிகவும் உலகின் இருக்கும் கூட உலக என்ற தொலைவு தென்னிந்தியாவின் அளவில்	மேலே குறிப்பிட்ட எண்களை ஓடும் சீனாவின் முதன்மையான ஆறுகளில் இருக்கின்றன அறிவித்தது மாற்றங்கள்	ஈடுபாடு மீதும் குறித்த கட்டுரையைத் தொடங்குங்கள் பிற கட்டுரைகளில் தேடிப்பாருங்கள் இலக்கிய கொண்டவர்	நதி உள்ள கோபாலகிருஷ்ணன் உங்களின் த.வி.யில் மொழி பங்கு வகிக்கிறது நன்றி சுந்தர்
நீங்கள் அடங்கும் ஸ்ரீநிவாசன் கருத்து உள்ளன ராமசாமி ராவ் என் அமைவது தற்போது	ஒன்று இரண்டு மூன்று நான்கு ஆகிய முக்கிய புகழ் பௌத்த ஆலயம் சிவாஜி கணேசன்		

2.4 Observation

(i) K Means Clustering Error



(ii) K Means Cluster Initialization

Due to random initialization, one cluster in all the 4 corpora is very big. Few clusters are empty also. But apart from that, words in the remaining clusters are pretty evenly distributed.

(iii) K Means Clustering Saturation

Reduction in error saturates after 4-5 iterations usually.