ICDCIT-2012

PROJECT INNOVATION CONTEST

Etola: A News Headline Clustering Tool

(PIC No: 11)

Nibir Nayan Bora

Project Guide:

Dr. Bhabani Shankar Prasad Mishra

KIIT University,

Bhubaneswar, Odisha

Outline

- Perform Cluster Analysis on news headlines.

- Create a dataset
- Propose a new technique
- Compare with traditional method
- Evaluate quality

Cluster Analysis

Assigning objects to groups (called *clusters*) in accordance with the general clustering rule, '*high intra-cluster similarity and low inter-cluster similarity*'

- Document clustering / Text categorization
- Unsupervised learning

Why News Headlines?

How are news headlines different?

- Small size
- Less dimensions
- Grammatical inconsistent

Benefit:

- Traditional clustering methods can be modified accordingly.

The Dataset

- Extracted from the Reuters corpus (90 category split)

- No. of documents: 343
- No. of classes (topics): 26
- e.g. *gold, coffee, jobs, retail, housing.*

- Each headline relates to a single topic.

The Dataset (*examples*)

Headline	Topic
U.S. cotton certificate expiration date extended	Cotton
China trying to increase cotton output, paper says	Cotton
January housing sales drop, realty group says	Housing
Quebec February housing starts fall	Housing
Pakistan not seen as major wheat exporter	Wheat
Weather hurting Yugoslav wheat - USDA Report	wheat

Preprocessing

- Unwanted characters
- Case normalization
- Tokenization
- Stopwords
- Stemming

- Vector Space Model

k-means

- Randomly select *k* initial centroids
- **repeat**
  - Assign each document to the closest centroid.
  - Recalculate cluster centroids.
- **until** clusters do not change

Frequent term

- Create a list of all terms in corpus and arrange them in order of their frequencies.
- **for each** term
  - **if** the term is present more often in unclustered documents, **then**

    Form a cluster by collecting all documents containing the term.
- Apply refinement

Refinement

- For clusters with single document, mark document as unclustered.
- Assign each unclustered document to its closest cluster.
- Apply *k-means* clustering to these clusters.

Frequent term *plus*

Assumption:

- The lesser the number of terms in a document, the easier it is to identify the topic term.
- A document contains only one topic term.

What do we do?

- Arrange the documents in reverse order of number of terms.
- Try to find a topic term in each.

Frequent term *plus* (*algorithm*)

- Create a list of all terms in corpus and their frequencies.
- Arrange the documents in ascending order of number of terms.
- **for each** document
  - Starting with term with highest frequency in corpus, try to find a topic term.
  - **if** the term is present more often in unclustered documents, **then**

    Form a cluster by collecting all documents containing the term.
- Apply refinement

Frequent nouns

- Similar to Frequent term clustering.
- In preprocessing, all non-noun features are removed.

Frequent nouns *plus*

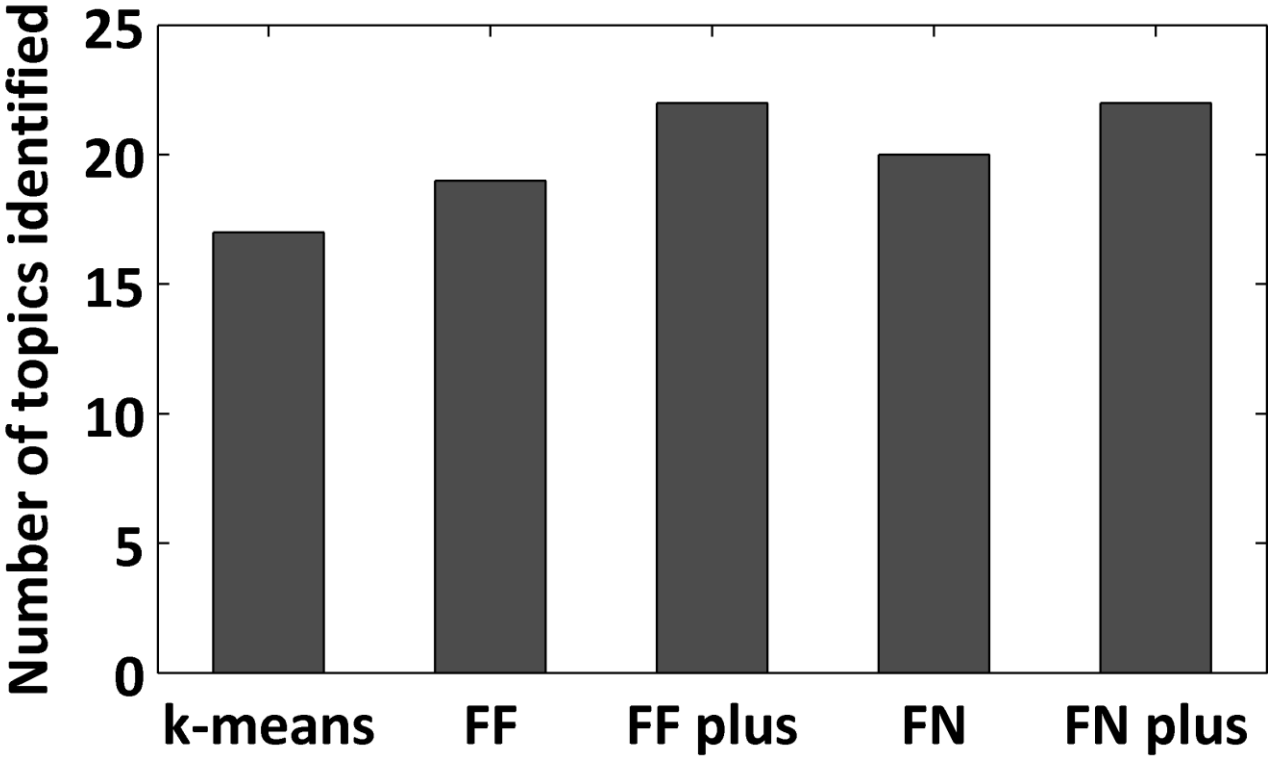
- Similar to Frequent term *plus* clustering.
- In preprocessing, all non-noun features are removed.

Results

Legend:

FF – Frequent feature

FN – Frequent nouns



	k-means	FF	FF *plus*	FN	FN *plus*
No. of clusters generated	26	29	34	28	33
Correct clusters	17	19	**22**	20	**22**
Incorrect clusters	9	10	**12**	8	11
No. of documents correctly clustered	238	226	255	241	**272**

Applications

- Create news categories by clustering news from various sources.
  - Like Google News

Future Work

- Cluster descriptor
- Use of synonyms