Data Mining Project Proposal

Group Doreamon:
Abdul Gafar Manuel Meque
Chih-Ming Chen
Sachit Mahajan

Problem Description

- What Knowledge?

  The relationship between **post emotion/polarity** and **its words** under different subreddit.

- Given a Post *p* with upvotes *u* and downvotes *d*
  - Can we predict the **opinion polarity** (*negative* | *positive*) based on *u* & *d?*
    - P(p=negative|u,d)
    - P(p=positive|u,d)
  - Can we Predict the ***u,d* distribution** based on *p?*
    - P(u,d|p)

Problem Description

- What Knowledge?

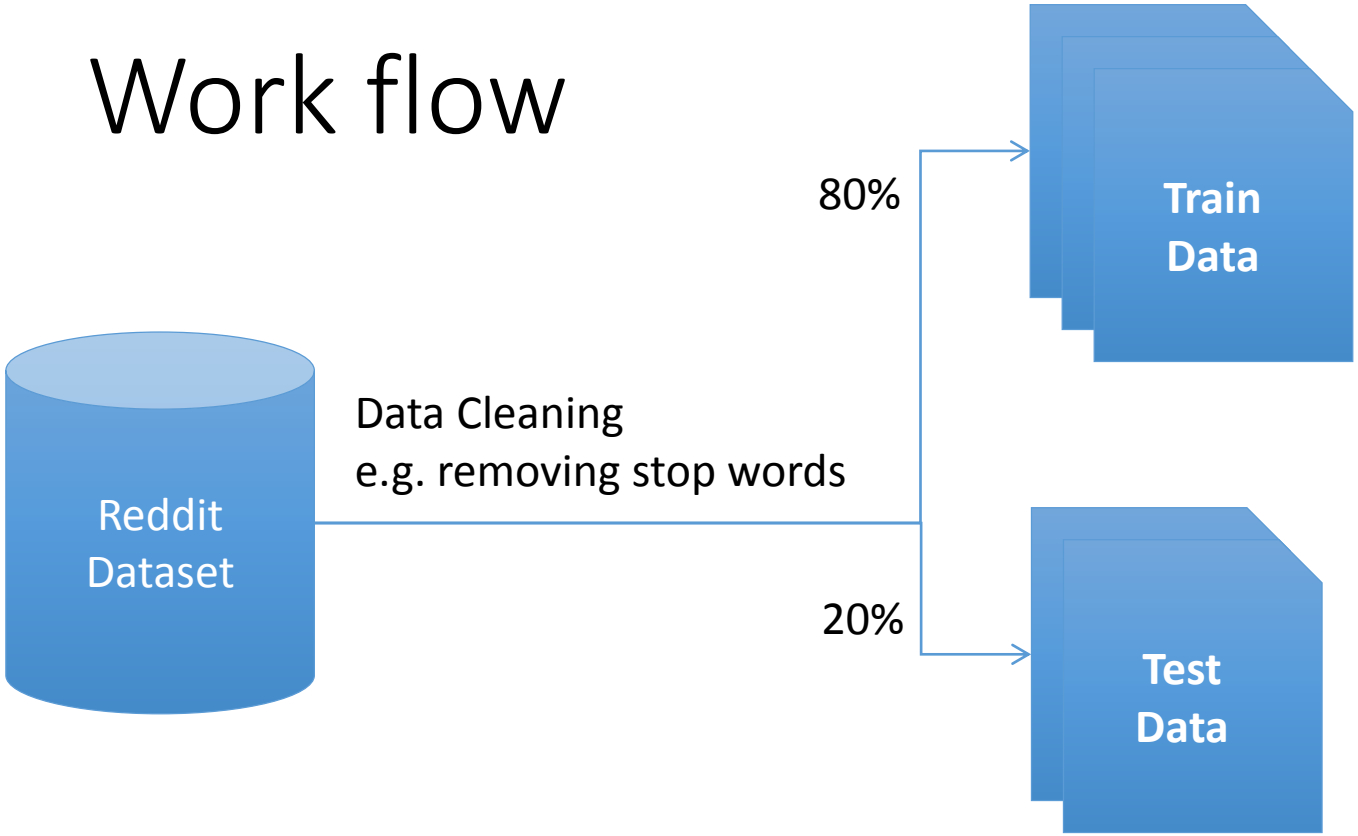
The relationship between **a word** and **its polarity** under different subreddit.

- Given a Post *p* with upvotes *u* and downvotes *d*
  - Can we predict the **opinion polarity** (*negative* | *positive*) based on *u* & *d?*
    - P(p=negative|u,d) $P(p_{pol} = neg|u, d)$
    - P(p=positive|u,d) $P(p_{pol} = pos|u, d)$
  - Can we Predict the ***u,d* distribution** based on *p?*
    - P(u,d|p) $P(d, u|p)$

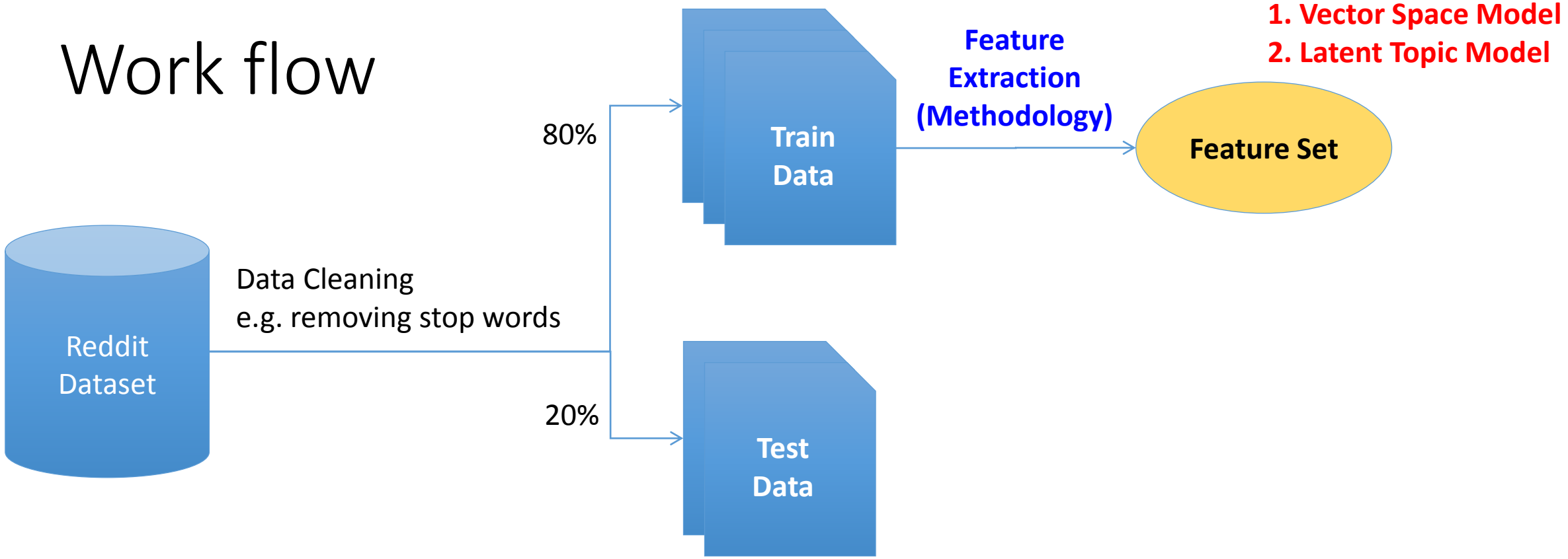
Work flow

Data Cleaning
e.g. removing stop words

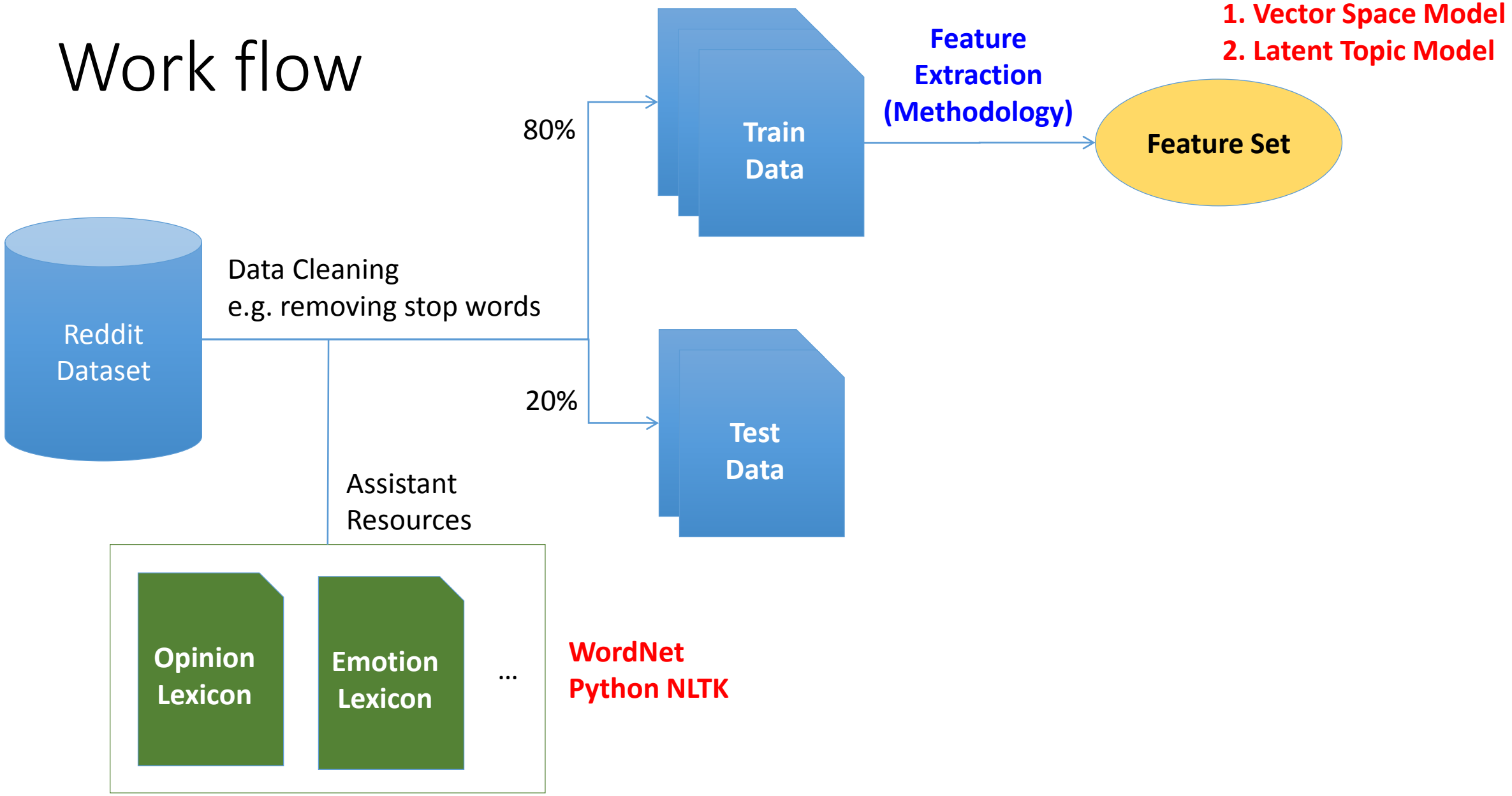
Work flow


Reddit
Dataset
Data Cleaning
e.g. removing stop words
80%
20%
Train
Data
Test
Data

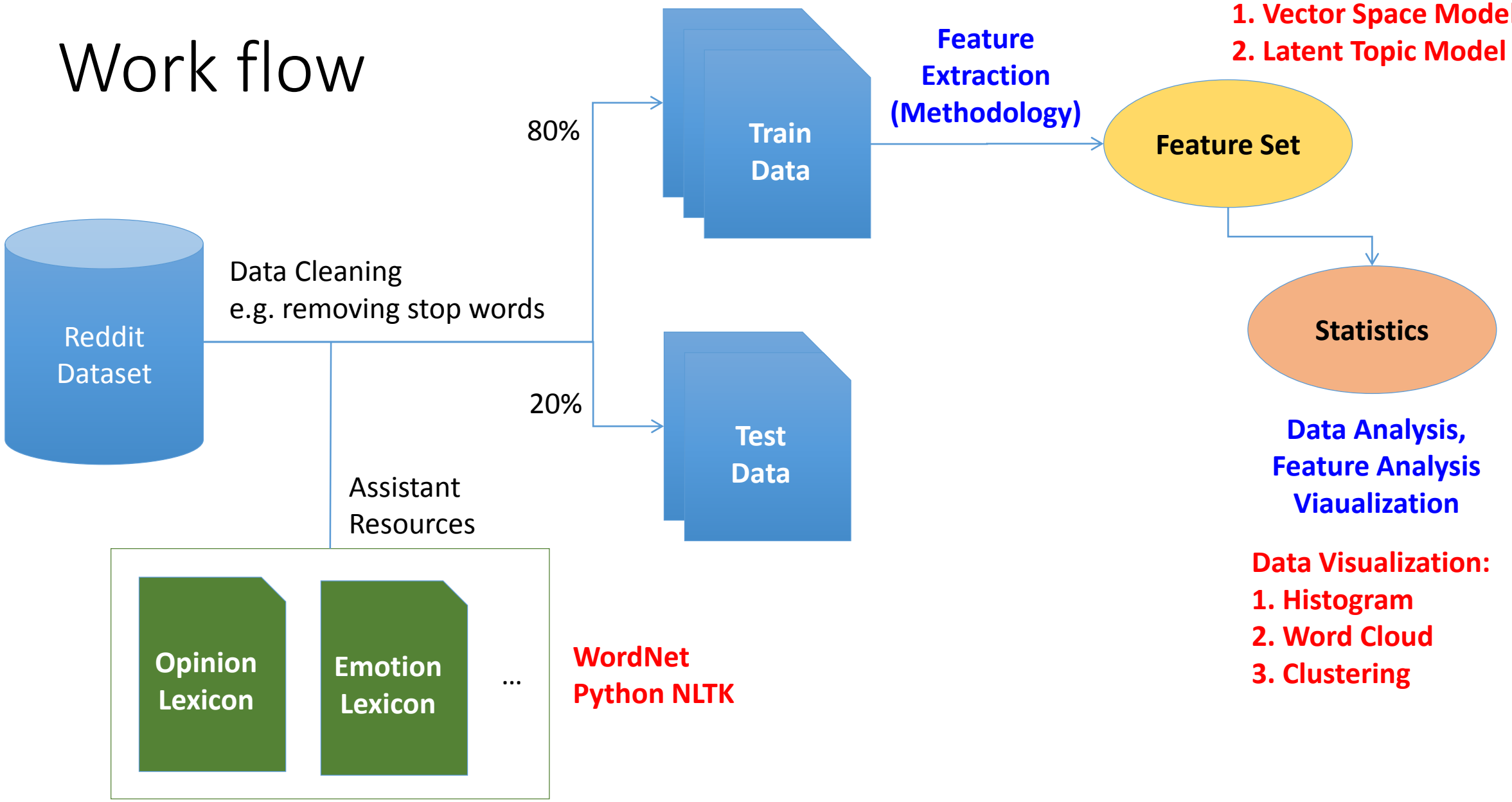
Work flow


Extracted features:
1. Vector Space Model
2. Latent Topic Model
Feature Extraction (Methodology)
Feature Set
Train Data
80%
Reddit Dataset
Data Cleaning
e.g. removing stop words
20%
Test Data

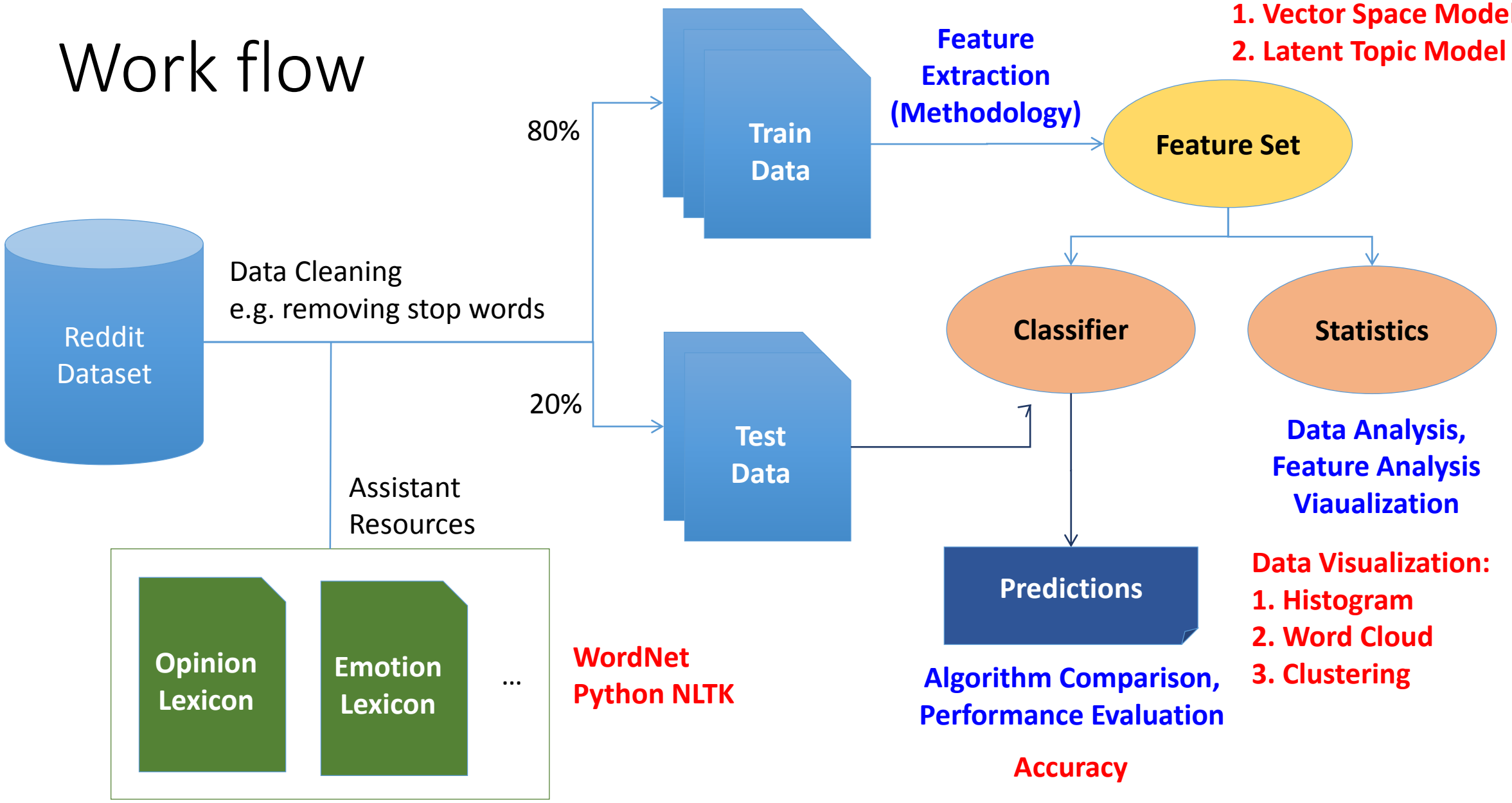
Work flow


Extracted features:
1. Vector Space Model
2. Latent Topic Model
Feature Extraction (Methodology)
Feature Set
Train Data
80%
Reddit Dataset
Data Cleaning
e.g. removing stop words
20%
Test Data
Assistant Resources
Opinion Lexicon
Emotion Lexicon
...
WordNet
Python NLTK

Work flow


Reddit Dataset
Data Cleaning
e.g. removing stop words
80%
20%
Train Data
Test Data
Feature Extraction (Methodology)
Feature Set
Statistics
Extracted features:
1. Vector Space Model
2. Latent Topic Model
Data Analysis,
Feature Analysis
Viaualization
Data Visualization:
1. Histogram
2. Word Cloud
3. Clustering
Assistant Resources
Opinion Lexicon
Emotion Lexicon
...
WordNet
Python NLTK

Work flow


Extracted features:
1. Vector Space Model
2. Latent Topic Model
Feature Extraction (Methodology)
80%
Train Data
Feature Set
Reddit Dataset
Data Cleaning
e.g. removing stop words
Classifier
Statistics
20%
Test Data
Data Analysis,
Feature Analysis
Viaualization
Assistant Resources
Predictions
Data Visualization:
1. Histogram
2. Word Cloud
3. Clustering
Opinion Lexicon
Emotion Lexicon
...
WordNet
Python NLTK
Algorithm Comparison,
Performance Evaluation
Accuracy

A Sketch

In a subreddit, given a post

Upvotes: 10

Downvotes: 1

Joseph Levy was preparing for a season of scientific research in Antarctica last week when he got the call: Stand down.

Dr. Levy, a research associate at the University of Texas at Austin's Institute for Geophysics, is studying the climate history of the dry valleys of Antarctica by analyzing buried ice sheets that have been frozen since the last ice age and are beginning to thaw.

The research season in Antarctica typically starts around now, when things warm up enough to be merely frigid and scientists from around the world flock far south to conduct studies that affect our understanding of climate change, volcanoes, the family life of Weddell seals and much more. But with the United States government partly shut down, federally financed research has come to a halt for Dr. Levy and hundreds of other Americans. Even if a budget deal is struck, these scientists will have less time on the ice, and some will lose a full year's worth of work as the narrow window of productive time closes.

"It's like a biography of the earth with a couple of pages in the middle torn out," Dr. Levy said. "Nature will have taken its course, and we will have not been there to see it."

The shutdown in Washington is being felt acutely at the ends of the earth. Some 3,000 Americans work through the Antarctic summer, including scientists and support staff

Positive?

Negative?

Neutral?