Feature Extraction from Financial News Corpora for Stock Prediction using Deep Learning

Natural Language Processing (CS-421)

University of Illinois at Chicago

Harish Naik

Sarit Adhikari

# Document Revision History

|  |  |  |
| --- | --- | --- |
| Date | Modified By | Remarks |
| 27th April 2017 | Sarit Adhikari | Initial draft |
|  |  |  |

Contents

[Document Revision History 2](#_Toc481077322)

[Introduction 4](#_Toc481077323)

[Goal 4](#_Toc481077324)

[Dataset 4](#_Toc481077325)

[Implementation 5](#_Toc481077326)

[Text Preprocessing 5](#_Toc481077327)

[Bag of words 6](#_Toc481077328)

# Introduction

## Goal

Our goal is to extract fixed-size features using word-embedding techniques suitable to DNNs from financial news corpora.

## Dataset

News articles from Bloomberg and Reuters from 2006/10/20 to 2013/11/20

# Implementation

## Text Preprocessing

Split financial articles into sentences

Keep sentences that mention at least one stock name or one public company

Each sentence is labelled by the publication date of the original article and the mentioned stock name

Group these sentences by the publication dates and the underlying stock names to form the samples

## Bag of words

We have used word2vec from genism library to generate bag of words from news corpora.

We have used seven seed words "surge", "rise", "shrink", "jump", "drop", "fall", "plunge", "gain", "slump"

Similar words are generated from these seed words recursively.

Following screenshot shows the snapshot of generation procedure.



These keywords will be the feature attributes for the sample. The feature values will be the tfidf score for a document (or sample)



