白皮书

## 网页内容抽取算法

给定任意的url，可以抽取网页的正文、关键词等信息

1. 通过jsoup获得html的DOM

## 基于fastText的网页安全性检测

web page content extraction algorithm

网页正文和关键词提取算法

1、通过jsoup获取网页的html

网页安全性检测算法

基于敏感词的网页安全检测

基于FastText的网页安全分类算法

FastText

FastText is a library created by the Facebook Research Team for efficient learning of word representations and text classification, we can use this for .

First, we have trained.

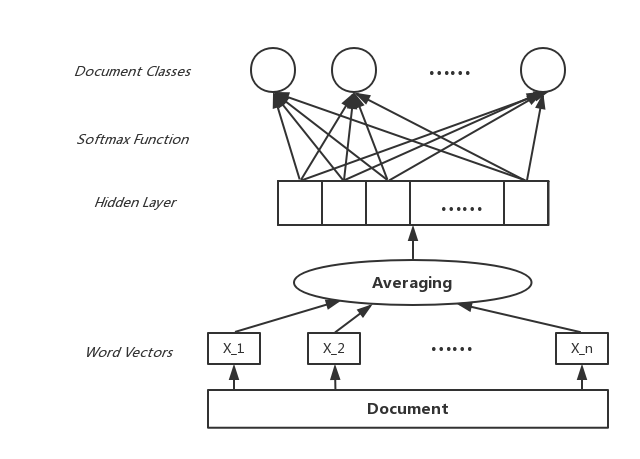
We trains the Word2Vec model from content of millions web page

without us

In unsupervised mode, fastText trains by sliding a window over the input text and either learning the centre word from the remaining context

The fastText classifier is a Linear Bag of Words (LBoW) classifier, which uses linear technique both for combining the word vectors into the vector representing the document, and for computing the classification criterion.

The structure for fastText classifier is shown in Fig.1.



First, the classifier computes a document vector y. This vector y is a linear bag of words of the document. It is computed by averaging all document word vectors x\_i

Here, n is the number of N-gram features x\_1,x\_2,x\_n.

the loss function is

fasttext_loss_function

其中，x\_n，y\_n是训练样本，矩阵参数A是基于word的look-up table，Ax\_n矩阵运算的数学意义是将word的embedding向量找到后取平均，得到hidden向量。矩阵参数B是函数f的参数，函数f是一个多分类问题。

how to train

At the beginning, We have collected training samples from

will use fastText to generate the model for representing words by vectors.

网页数据清洗

爬取网页内容后，获取<meta>中的title、keywords、description数据，

分词，停用词，词性标注，屏蔽某些词性

得到候选词，其中title和keyword中得到的词权重最高。

LDA提取

url分类

fasttext

收集

url提取关键词

Word Segmenter

## Content Extraction of Web Page

## Web Page Safety Detection Based on FastText