FeatureMagic 0.0.1

# FeatureMagic简介

FeatureMagic是一个基于Java实现的自然语言处理工具，该工具从cohesion、language和readability三个角度对文本进行特征分析和提取。目前仅仅针对英文。***但是最终的版本是语言无关，或者说可以通过修改配置文件来对不同语言的文本进行特征抽取。***

## 开发人员

FeatureMagic特征提取库是由北京大学2013级硕士王志伟在俞敬松副教授的指导下完成的。

# FeatureMagic架构

# FeatureMagic核心算法

## 3.1 Descriptive

## 3.2 Text Easability Principal Component Scores

## 3.3 Referential Cohesion

## 3.4 LSA

## 3.5 Lexical Diversity

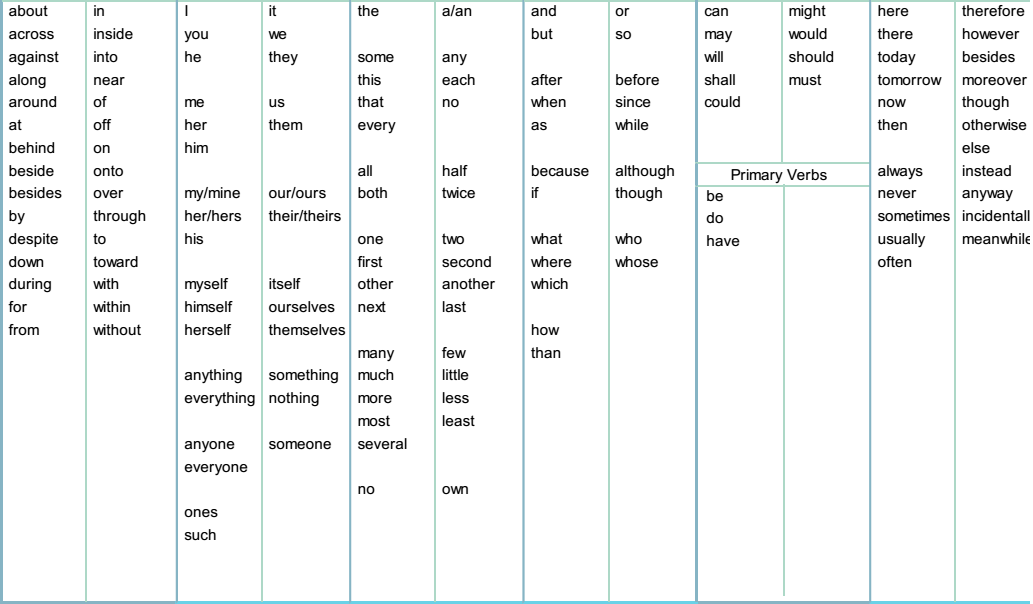
### 3.5.1 TTR for Content Words

TTR for Content Words是指求一篇自然语言文本实词的类型单词数量比值。在详细介绍该算法之前，首先要介绍几个术语和本算法采用的必须数据。

[1] TTR：Type-token ratio，即自然语言文本单词类型数量和单词数量的比值。

[2] Content Words: 表示英语中的实词，一般情况下英语语言中的单词主要分为实词和虚词两类，即Content Words和Function Words。或者进一步地说，英语中的单词要么是实词要么是虚词。在计算实词的TTR值时，需要过滤掉虚词，由于实词数量庞大而虚词数量有限，因此选择过滤数量有限的虚词，这时需要一个必须数据——虚词列表。

虚词列表由各种版本，每种版本都有细微的差异，主要的差异来源是由于虚词列表的应用领域和划分角度不同造成的。FeatureMagic采用的虚词列表是来自于[1]，列表如下：



标点符号列表。FeatureMagic的设计是与语言无关的，所以用户可以通过修改配置文件和设置对应文本自然语言的标点符号，在本说明中是针对英语来说，标点符号列表为

.?!,;:-'()[]

### 3.5.2 MTLD

## 3.6 Connectives

## 3.7 Situation Model

## 3.8 Syntactic Complexity

## 3.9 Syntactic Pattern Density

## 3.10 Word Information

# FeatureMagic使用指南

## 4.1 安装步骤

## 4.2 命令行工具调用

## 4.3 API调用

### 4.3.1 Java接口

### 4.3.3 Python接口

*后期实现。*

## 4.4 GUI调用

# 扩展FeatureMagic

## 5.1 增加新特征

在项目引用了FeatureMagic库之后，只需要继承Feature抽象类和Extractor接口就可以增加新的特征类，在这个新的特征类中，可以基于Text类计算出的文本的基础数据来计算新的特征。当Text类中计算的基础数据不能够满足计算新的特征需要时，也可以继承Text类来进行新的基础数据的计算。

## 5.2 增加新语言支持

FeatureMagic特征抽取库在上层是与文本的自然语言无关的，但是在底层处理自然语言文本时调用了第三方自然语言处理库，而这些第三方自然语言处理库是与具体的自然语言相关的。为了让FeatureMagic能够根据用户的需要来定制，FeatureMagic在设计时并没有与第三方进行绑死，所以用户只需要通过修改相应的配置文件来定制FeatureMagic调用的第三方库和指定将要处理文本的自然语言，这样就可以使用FeatureMagic来对任意自然语言的文本进行特征提取。

# FeatureMagic引用的第三方库

FeatureMagic特征提取库的实现，引用了部分第三方自然语言处理工具包，这些工具包如下：

[1] OpenNLP 3.0.3

[2] StanfordNLP 3.2.0

[3] LingPipe 4.1.0

# 参考文献

[1] <http://www.flesl.net/Vocabulary/Single-word_Lists/function_word_list.php>。