Atitit 自动摘要法与 **textrank算法**

自动文摘的方法。自动文摘（[Automatic Summarization](https://en.wikipedia.org/wiki/Automatic_summarization" \t "https://www.cnblogs.com/clover-siyecao/p/_blank)）的方法主要有两种：Extraction和Abstraction。其中Extraction是抽取式自动文摘方法，通过提取文档中已存在的关键词，句子形成摘要；Abstraction是生成式自动文摘方法，通过建立抽象的 语意表示，使用自然语言生成技术，形成摘要。由于生成式自动摘要方法需要复杂的自然语言理解和生成技术支持，应用领域受限。所以本人学习的也是抽取式的自动文摘方法。

# 目前主要方法有：

## 基于统计：统计词频，位置等信息，计算句子权值，再简选取权值高的句子作为文摘，特点：简单易用，但对词句的使用大多仅停留在表面信息。

## 基于图模型：构建拓扑结构图，对词句进行排序。例如，TextRank/LexRank

## 基于潜在语义：使用主题模型，挖掘词句隐藏信息。例如，采用LDA，HMM

## 基于整数规划：将文摘问题转为整数线性规划，求全局最优解。

# ****textrank算法****

**TextRank算法基于PageRank，用于为文本生成关键字和摘要。**

**TextRank**

[TextRank](http://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf" \t "https://www.cnblogs.com/clover-siyecao/p/_blank)算法是一种用于文本的基于图的排序算法。其基本思想来源于谷歌的 [PageRank](http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf" \t "https://www.cnblogs.com/clover-siyecao/p/_blank)算法, 通过把文本分割成若干组成单元(单词、句子)并建立图模型, 利用投票机制对文本中的重要成分进行排序, 仅利用单篇文档本身的信息即可实现关键词提取、文摘。和 LDA、HMM 等模型不同, TextRank不需要事先对多篇文档进行学习训练, 因其简洁有效而得到广泛应用。