**1. 基本文本处理技能**

1.1 分词的概念（分词的正向最大、逆向最大、双向最大匹配法）；

1.2 词、字符频率统计；（可以使用Python中的collections.Counter模块，也可以自己寻找其他好用的库）

**2. 概念**

2.1 语言模型中unigram、bigram、trigram的概念；

2.2 unigram、bigram频率统计；（可以使用Python中的collections.Counter模块，也可以自己寻找其他好用的库）

**3. 文本矩阵化：要求采用词袋模型且是词级别的矩阵化**

步骤有：

3.1 分词（可采用结巴分词来进行分词操作，其他库也可以）；

3.2 去停用词；构造词表。

3.3 每篇文档的向量化。

1.1 分词的概念（分词的正向最大、逆向最大、双向最大匹配法）；

百度百科定义：中文分词(Chinese Word Segmentation) 指的是将一个汉字序列切分成一个个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。维基百科定义：Word segmentation is the problem of dividing a string of written language into its component words.

简单的讲，中文分词就是让计算机系统在汉语文本中的词与词之间自动加上空格或其他边界标记。

汉语分词难点：分词规范、歧义切分、未登录词识别。

分词规范：词这个概念一直是汉语语言学界纠缠不清又挥之不去的问题，也就是说，对于词的抽象定义(词是什么)和词的具体界定(什么是词)迄今拿不出一个公认的、具有权威性的词表来。

歧义切分：切分歧义是汉语分词研究中一个大问题，因为歧义字段在汉语文本中大量存在。处理这类问题可能需要进行复杂的上下文语义分析，甚至韵律分析(语气、重音、停顿等)。

未登录词识别：未登录词又叫生词，一般有二种解释：第一种指的是已有的词表中没有收录的词；第二种指的是已有的训练语料中未曾出现过的词。在第二种解释下，又称之为集外词(out of vocabulary，OOV)。

中文分词基本方法：

三大类：基于词典的方法、基于理解的分词方法和基于统计的分词方法。

基于词典的方法(又称机械分词方法)：

这种方法本质上就是字符串匹配的方法，将一串文本中的文字片段和已有的词典进行匹配，如果匹配到，则此文字片段就作为一个分词结果。但是基于词典的机械切分会遇到多种问题，最为常见的包括歧义切分问题和未登录词问题

常用的字符串匹配方法有如下几种：

正向最大匹配法（从左到右的方向）；

逆向最大匹配法（从右到左的方向）；

最少切分（每一句中切出的词数最小）

双向最大匹配（进行从左到右、从右到左两次扫描）

这类算法的优点是速度快，时间复杂度可以保持在O（n）,实现简单，效果尚可；但对歧义和未登录词处理效果不佳。

1.正向最大匹配分词

正向的就是从前往后比较，要设置一个最长词的长度maxLength，就是最长能识别这么长的词

双向最大匹配法是将正向最大匹配法得到的分词结果和逆向最大匹配法的到的结果进行比较，从而决定正确的分词方法。据SunM.S. 和 Benjamin K.T.（1995）的研究表明，中文中90.0％左右的句子，正向最大匹配法和逆向最大匹配法完全重合且正确，只有大概9.0％的句子两种切分方法得到的结果不一样，但其中必有一个是正确的（歧义检测成功），只有不到1.0％的句子，或者正向最大匹配法和逆向最大匹配法的切分虽重合却是错的，或者正向最大匹配法和逆向最大匹配法切分不同但两个都不对（歧义检测失败）。这正是双向最大匹配法在实用中文信息处理系统中得以广泛使用的原因所在。

实际的分词过程中可以加入一些启发式的规则来对分词结果进行进一步消歧的。

启发式规则：

1.如果正反向分词结果词数不同，则取分词数量较少的那个。

2.如果分词结果词数相同

a.分词结果相同，就说明没有歧义，可返回任意一个。

b.分词结果不同，返回其中单字较少的那个。

中文分词原则

分词粒度以大为最佳（最大化平均词语长度）

分词的结果颗粒度（单个词所包含的字符长度）越大越好。

同样是 北京大学 可以分成 北京 / 大学 或者 北京大学，则后者更优，后者包含的信息更加特定和明确。

非词典词越少越佳，单字字典词数越少越佳

非词典词的出现说明出现了 OOV (Out Of Vocabulary) 问题，字典足够大的情况下出现 OOV， 说明分词结果不佳。类似的，分词结果大量出现单个字，也是暗示分词效果不佳。

比如 技术和服务 可以被分成 技术 / 和 / 服务 或者 技术 / 和服 / 务，后者中的 务 就是一个 OOV，因为中文中 务 不能单独成词

最小化词语长度的变化率

同样是 研究生命起源 可以被分成 研究生 / 命 / 起源 和 研究 / 生命 / 起源，后者 词语长度的变化率 最小，因此是更好的分词结果。

2.1 语言模型中unigram、bigram、trigram的概念；

unigram 一元分词，把句子分成一个一个的汉字

bigram 二元分词，把句子从头到尾每两个字组成一个词语

trigram 三元分词，把句子从头到尾每三个字组成一个词语.

* 1. unigram、bigram频率统计；

unigram(这和栗子不是很好频率都是1)

文本矩阵化：要求采用词袋模型且是词级别的矩阵化

步骤有：3.1 分词（可采用结巴分词来进行分词操作，其他库也可以）；  
3.2 去停用词；构造词表。  
3.3 每篇文档的向量化。