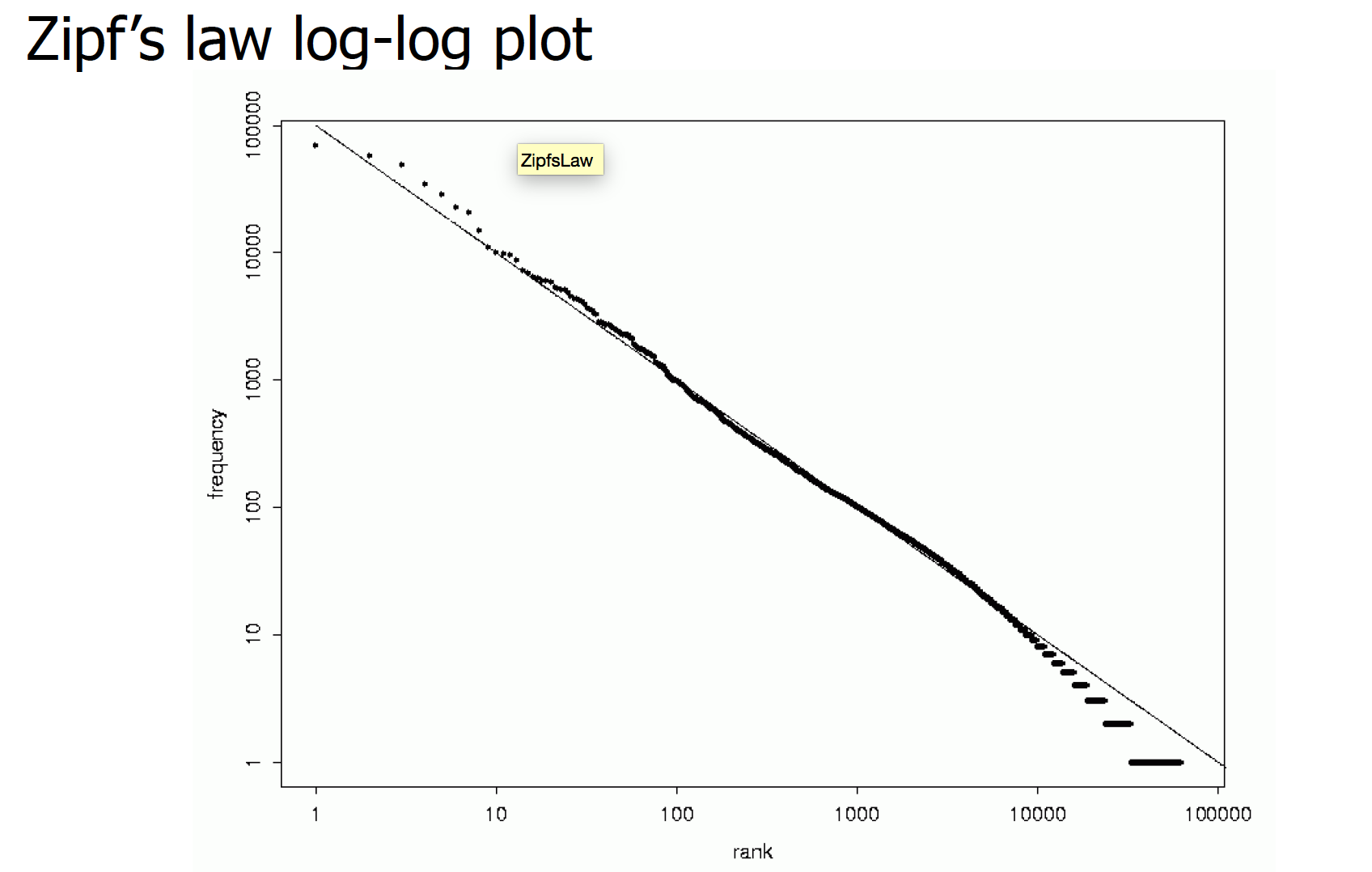
使用**给定的北京大学《人民日报》标注语料库**，完成Zipf定律曲线生成和基于n-gram的句子概率计算。具体要求如下：

1 **Zipf定律曲线生成**

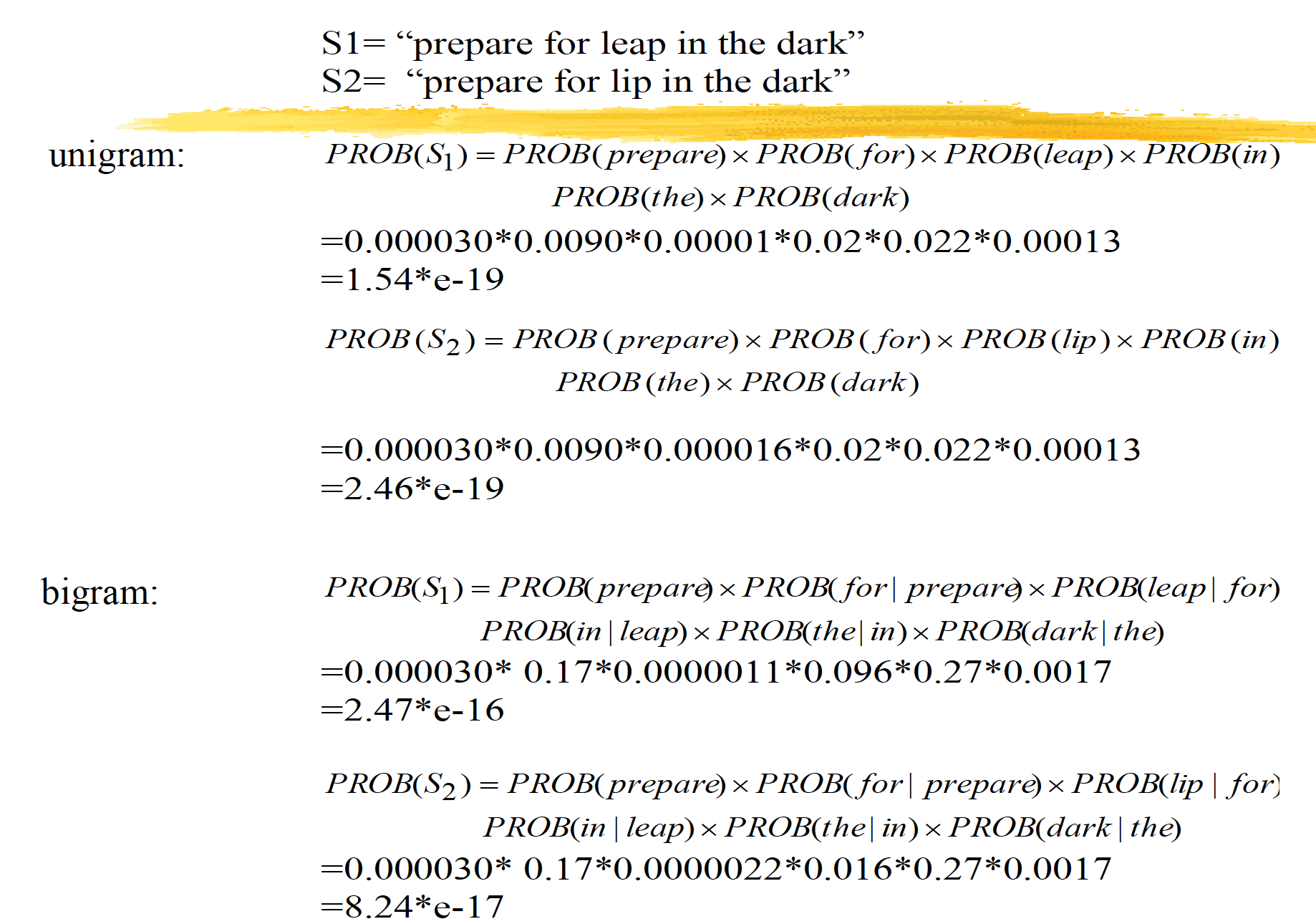
1. 统计已人工分词的语料库中所有词的**词频**，并按词频由高到低进行排序，生成文件“1.txt”。
2. 每行输出一个词的信息，先输出词，再输出该词的频度，中间用tab分隔。
3. 基于该词频表，利用Excel生成Zipf定律的曲线图，生成文件“2.jpeg”。
4. 该曲线图的横坐标为rank（取log），纵坐标为frequency（取log），请参考lesson4part1课件第28页。



1. 对得到的曲线进行必要的观察和分析。

2**基于n-gram的句子概率计算**

1. 对北京大学《人民日报》标注语料库适当处理，如：在每一句开头增加一个句子起始符<BOS>, 在其结尾处增加一个句子结束符<EOS>。 （可以认为这两个特殊字符是“虚拟”地加上去的）
2. 分别使用**unigram**和**bigram**计算以下两个句子的概率（参考lesson4part1课件第16页）：



* 扶贫 开发 工作 取得 很 大 成绩 （句子1）
* 扶贫 开发 工作 得到 很 大 成绩 （句子2）
* <BOS> 扶贫 开发 工作 取得 很 大 成绩 （<句子1a）
* <BOS> 扶贫 开发 工作 得到 很 大 成绩 （<句子2a）
* <BOS> 扶贫 开发 工作 取得 很 大 成绩 <EOS> （句子1b）
* <BOS> 扶贫 开发 工作 得到 很 大 成绩 <EOS> （句子2b）

1. 生成文件“3.txt”，格式如下：

* 第一行输出句子1的unigram句子概率，并分别输出每个unigram的概率。
* 第二行输出句子1的bigram句子概率，并分别输出每个bigram的条件概率。
* 第三行输出句子1a的bigram句子概率，并补充输出与<BOS>相关的bigram条件概率。
* 第四行输出句子1b的bigram句子概率，并补充输出与<BOS>和<EOS>相关的bigram条件概率。
* 对句子2同上依次处理。
* 行内以tab分隔，概率输出取log（10为底）结果，小数点后保留6位。

1. 对得到的上述结果进行必要的**观察和分析**。

每位同学在网络学堂上提交“1.txt”、“2.jpeg”和“3.txt”及**相关观察和分析**。