1. 什么是文本分类？

答：文本分类用电脑对文本集(或其他实体或物件)按照一定的分类体系或标准进行自动分类标记，属于一种基于分类体系的自动分类，是朴素贝叶斯分类方法。基于分类体系的自动分类，基于资讯过滤和用户兴趣(Profiles)的自动分类所谓分类体系就是针对词的统计来分类关键字分类，现在的全文检索词的正确切分不易分辨(白痴造句法)学习人类对文本分类的知识和策略，从人对文本和类别之间相关性判断来学习文件用字和标记类别之间的关联。

2.文本分类有什么应用场景？

答：1. 新闻主题分类（文章分类）：根据文章内容（或者结合标题）给新闻等其他文章一个类别，比如财经、体育、军事、明星等等。一般在新闻资讯方面使用比较多。  
2. 情感分析：两类（正面、负面）或者多类（例如，生气、高兴、悲伤等等），其实还是有三类（正、负、中性），不过和两类的处理方法会有些许区别。一般在影评（比如豆瓣、淘票票）、商品评价（比如淘宝、京东的商品评价）等对商品和服务的评价方面应用比较多。  
3. 舆情分析：和情感分类类似，更多的是两分类，政府或者金融机构用的比较多。

1. 什么是文本聚类？

答：文本聚类(Text clustering)文档聚类主要是依据著名的聚类假设:同类的文档相似度较大，而不同类的文档相似度较小。作为一种无监督的机器学习方法，聚类由于不需要训练过程，以及不需要预先对文档手工标注类别，因此具有一定的灵活性和较高的自动化处理能力，已经成为对文本信息进行有效地组织、摘要和导航的重要手段，为越来越多的研究人员所关注。

1. 文本聚类有什么应用场景？

答：1.文本聚类可以作为多文档自动文摘等自然语言处理应用的预处理步骤，比较典型的例子是哥伦比亚大学开发的多文档文摘系统Newsblaster。Newsblaster将每天发生的重要新闻文本进行聚类处理，并对同主题文档进行冗余消除、信息融合、文本生成等处理，从而生成一篇简明扼要的摘要文档;

2.对搜索引擎返回的结果进行聚类，使用户迅速定位到所需要的信息。Hua-Jun Zeng等人提出了对搜索引擎返回的结果进行聚类的学习算法。比较典型的系统则有vivisimo和infonetware等。系统允许用户输入检索关键词，而后对检索到的文档进行聚类处理，并输出各个不同类别的简要描述，从而可以缩小检索的范围，用户只需关注比较有希望的主题。另外这种方法也可以为用户二次检索提供线索;

3.对用户感兴趣的文档(如用户浏览器cache中的网页)聚类，从而发现用户的兴趣模式并用于信息过滤和信息主动推荐等服务。

4.聚类技术还可以用来改善文本分类的结果，如俄亥俄州立大学的Y.C. Fang, S. Parthasarathy和F. Schwartz等人的工作。

5.数字图书馆服务。通过SOM神经网络等方法，可以将高维空间的文档拓扑保序地映射到二维空间，使得聚类结果可视化和便于理解，如SOMlib[ ]系统;

6.文档集合的自动整理。如Scatter/Gather[ ]是一个基于聚类的文档浏览系统。而微软的Ji-Rong Wen等人则利用聚类技术对用户提出的查询记录进行聚类，并利用结果更新搜索引擎网站的FAQ。