大数据试卷

Q1.

A. 描述大数据分析中的 5 个 V
Volume, variety, velocity, veracity, value　数量、种类、速度、准确性、价值

B. 描述数据科学和商业智能时间的区别?
商业智能主要提供一些事后洞察和意见，通常用于解释事件的“时间”和“地点”。
数据科学主要以更具前瞻性和探索性的方式使用分类数据，侧重于分析现状，为未来决策提供数据参考。
Business intelligence mainly provides some after the fact **insights and opinions**, which are generally used to explain the "time" and "place" of events.
Data science mainly to classify data in a more forward-looking and exploratory way, focusing on analyzing the current situation and providing data reference for future decision-making.

C. 数据分析的六个阶段?

Discovery 发现

明确项目的可用资源，包括数据，数据模型，工具，系统，开发人员。
为大数据项目收集足够的多的数据。
Identify the available resources for the project, including data, data models, tools, systems, and developers.
Collect enough data for big data projects.

Data preparation 资料准备

此步骤主要包括数据清洗、数据转化、数据抽取、数据合并、数据计算等方法，目的是将原始数据加工成数据分析所需要的格式，例如：JSON、CSV 等等。
也可以利用数据可视化工具进行数据观察。
This step mainly includes data cleaning, data **conversion, data extraction,** data **merging,** data calculation and other methods. The purpose of this step is to process the original data into the **format required** for data analysis, such as JSON, CSV, txt,etc.
Data visualization tools can also be used for data observation.

Model planning 模型规划

根据数据特征选取适合数据的候选模型，例如：聚类、分类、关联等等。
了解和探索变量之间的关系，捕捉基本的预测因素和变量。
考虑数据是适用结构化、非结构化还是混合方法。
According to the **characteristics** of the data, we select suitable candidate models, such as clustering, classification, association and so on.
Understand and explore the relationship between variables, capture the basic predictors and variables.
Consider whether data is structured, unstructured, or hybrid.

Model building 搭建模型

将数据源分为训练集和测试集，并且反复迭代，达到训练效果均衡的目的。
实验中，应该记录模型的结果和逻辑和操作。
观察测试数据是否精准和有效，模型参数是否合理。
工具：python，R，matlab.....之类的

The data source is divided into training set and test set, and repeated iteration to achieve the goal of training effect balance.

In experiments, the results and logic and operations of the model should be recorded.

Observe whether the test data is accurate and effective, and whether the model **parameters** are reasonable.

Tools: Python, R, MATLAB, etc

Communication results 沟通结果

执行可靠的分析，寻求显示结果；对未来的工作提出建议，开始在实际生产环境中实施逻辑。

Perform reliable analysis, seek to display results, make suggestions for future work, and start to **implement logic** in actual production environment.

Operationalizec 操作化

将工作扩大到整个企业或者用户生态系统之前，设立一个试点项目，已受控方式部署工作；有效管理风险。

交付结果：项目介绍、项目演示、技术代码、技术规范

Set up a pilot project to deploy work before extending the work to the entire enterprise or user ecosystem; effectively manage risks.

Delivery results: project introduction, project demonstration, technical code, technical specification

Q2.

A. 解释为什么可视化在数据分析中很重要?

- 可视化提供简洁、整体的视图
- 可视化是初始数据探索的一个重要方面
- Visualization gives a clear, entire view
- Visualization is an important part at the initial data exploration

B. 那个 R 函数能够创建和加载 R 数据集?

Load file: read.csv()

C. 介绍分类中的假积极率（判断错的）和真积极率（判断对的）？

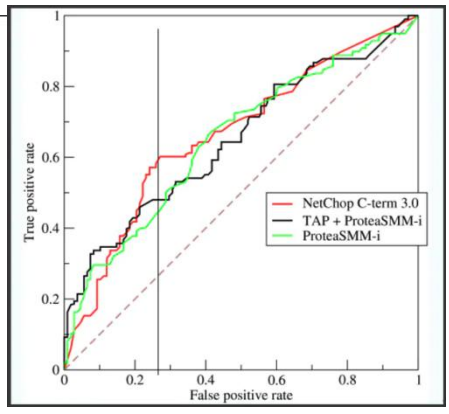
False positive rate(FPR) and true positive rate(TPR):

新的分类模型 performance 评判方法 New performance evaluation method for classification model

ROC 的全名叫做 Receiver Operating Characteristic，其主要分析工具是一个画在二维平面上的曲线——ROC curve。平面的横坐标是 false positive rate(FPR)，纵坐标是 true positive rate(TPR)

The full name of **ROC is receiver operating characteristic**. Its main analysis tool is a curve drawn on a two-dimensional plane - ROC curve. The abscissa of the plane is false positive rate (FPR), and the ordinate is true positive rate (TPR)

		真实值		总数
		p	n	
预测输出	p'	真阳性 (TP)	伪阳性 (FP)	P'
	n'	伪阴性 (FN)	真阴性 (TN)	N'
总数		P	N	

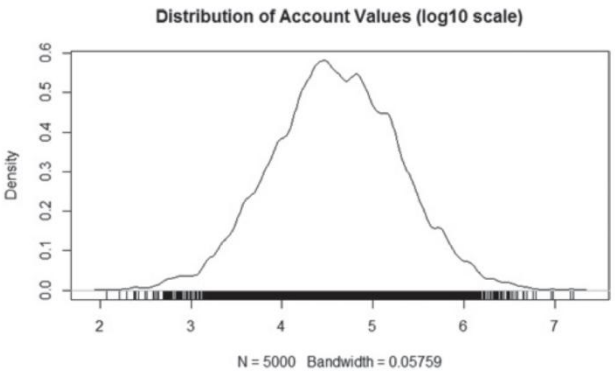


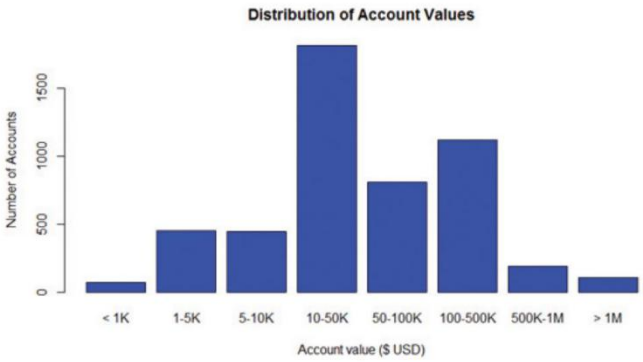
曲线越往右越好

TPR：在所有实际为阳性的样本中，被正确地判断为阳性之比率。TPR=TP/(TP+FN)

FPR：在所有实际为阴性的样本中，被错误地判断为阳性之比率。FPR=FP/(FP+TN)

TPR: the ratio of positive samples correctly judged to be positive. TPR=TP/(TP+FN)
FPR: the rate of false positives in all actual negative samples. FPR=FP/(FP+TN)
D.
Missed:上面的表不符合正态分布，应该改成以下图表
The above table does not conform to the **normal distribution**





Q3.

A. 解释假设测试的目的。假设你收到一个任务去识别一个新的计算模型程序在天气预测上是否有更好的精准度? 规定无效假设和可代替假设给这个任务 ？

假设检验的基本概念是提出假设，然后用数据进行检验。
原假设：新模型的精度与原模型一致。
替代假设：新模型的精度高于或者低于原模型。
The basic concept of **hypothesis** testing is to put forward hypotheses and test them with data.
Original hypothesis: the accuracy of the new model is consistent with that of the original model.
Alternative hypothesis: the accuracy of the new model is higher or lower than that of the original model.

B. 这是这几个概念 ？ Significance level 、 t-statistics: 、 p-value 、 Confidence interval:

Significance level：显著性水平
显著性水平：当原假设为真时，拒绝原假设的概率
the probability of rejecting the original hypothesis, when the null hypothesis is actually TRUE

t-statistics: T 统计
假设两个总体的分布具有相等但未知的方差
假设每个种群都是正态分布的
Suppose that the distributions of two populations have equal but unknown **variances**
Suppose that each population is normally distributed

p-value: p 值
p 值提供了观察到| T |≥T 的概率，前提是假设为真
p-value offers the probability of observing |T| ⩾ t given the null hypothesis is TRUE.

Confidence interval：**置信区间**
置信区间是指由样本统计量所构造的总体参数的估计区间
Confidence interval is the estimation interval of **population parameters** constructed by sample

statistics
置信区间展现的是这个参数的真实值有一定概率落在测量结果的周围的程度，其给出的是被测量参数的测量值的可信程度，即前面所要求的“一个概率”。

C. 描述方差分析的目标和过程？Goal and procedure of analysis of variance (ANOVA):
目标：用于样本均数差别的显著性检验。
Objective: to test the **significance of mean** difference between samples.
过程：
1.提出原假设
2.选择检验统计量：**方差分析**采用的检验统计量是 **F 统计量**，即 F 值检验。
=》
计算组间均值平方和
计算组内均值平方和
Process:
1. Put forward the original hypothesis
2. Select test statistic: the test statistic used in ANOVA is **F statistic**, that is, f-value test.
=》
Calculate the ***sum of squares of the mean*** values between groups
Calculate the sum of squares of the mean within the group

$$F = \frac{\text{组间均值平方和}}{\text{组内均值平方和}} = \frac{S_B^2}{S_W^2}$$

3.计算检验统计量的观测值和概率 P 值
4.给定显著性水平，并作出决策
3. Calculate the observed value and probability p value of the test statistic
4. Give significance level and make decision

Q4.

A. **描述 K-means 聚类的步骤**

1. 选择 k 的值和 k 的初始猜测值
2. 计算从每个数据点到每个中心的距离。将每个点指定给最近的中心线。
3. 更新每个簇的中心线
4. 重复步骤 2 和 3，直到收敛

1. Choose the value of k and the k initial guess for the **centriods**
2. Compute the distance from each data point to each centriod. Assign each point to the closest centriod.
3. Update the centriod of each cluster
4. Repeat Steps 2 and 3 until **convergence**

B. **描述 WWS 以及它在 k-means 聚类中的作用**

WSS 是什么？每个数据点和最近质心之间距离的平方和。

作用：WSS 是所有数据点与其最近质心之间距离的平方和。如果这些点相对接近各自的质心，则 WSS 将相对较小。因此，如果 K+1 聚类没有显著降低 K-meas 中的 WSS 值，那么添加一个聚类可能就不显著了。

WSS: Sum of the squares of the distances between each data point and the closest centroid.

WSS is the sum of squares of the distances between all data points and their nearest centroid. If these points are relatively close to their respective centroids, the WSS will be relatively small. Therefore, if K + 1 clustering does not significantly reduce the WSS value in k-means, then adding a cluster may not be significant.

C. 为什么我们通常运行多次 k-means 得到一个 K 值，k-means 中的说明方法用来记录次数 ？

对一个特定的 K 值运行多个 K 均值分析是非常重要的，以确保聚类结果具有总体最小 WSS。

It is very important to run multiple K-means analysis for a specific k-value to ensure that the clustering results have the **overall minimum WSS**. 使用了 nstart 函数来记录迭代次数。

D. 说出用于分类数据的聚类算法 ？

k-modes 算法： 是 K-Means 算法的一种扩展，适合于离散属性的数据集。采用简单匹配方法来度量分类型数据的相似度

K-Modes algorithm: it is an extension of K-means algorithm, which is suitable for data sets with **discrete** attributes. A simple matching method is used to measure the similarity of classified data。

Q5

A . 下面哪些项目的最小支持度大于 0.5 ？

{milk, egg, ham} , {milk, ham} , {egg, ham} , {milk, egg, ham ,butter} ,{milk, butter, ham}
{milk, ham} = 1, {egg, ham} = 3/5

可信度（置信度）：针对如{尿布}->{葡萄酒}这样的关联规则来定义的。计算为 支持度{尿布，葡萄酒}/支持度{尿布}，其中{尿布，葡萄酒}的支持度为 3/5，{尿布}的支持度为 4/5，所以“尿布->葡萄酒”的可行度为 3/4=0.75，这意味着尿布的记录中，我们的规则有 75%都适用。

B. 描述下列概念： 支持度、可信度、提升度、杠杆作用
1.支持度 (Support) ==» 支持度表示项集{X,Y}在总项集里出现的概率
2.置信度 (Confidence) ==» 即在含有 X 的项集中，含有 Y 的可能性
3.提升度 (Lift) ==» 提升度表示含有 X 的条件下，同时含有 Y 的概率 **与** 不含 X 的条件下却含 Y 的概率之比 ==> 意思就是说在含有 X 对含有 Y 的影响程度。
4.Leverage == > 测量 X 和 Y 出现在一起的概率与如果 X 和 Y 在统计上相互独立的话预期的差异

1. Support ==> the probability that the item set {x, y} appears in the total item set
2. Confidence ==>, that is, the possibility of containing y in the item set containing X
3. Lift ==> the ratio of the probability of both containing y with X and the probability of containing y without x == > which means the influence of X on the presence of Y.
4. Leverage ==> measures the difference between the probability of X and Y appearing together and the expected difference if x and y are statistically independent

1.支持度（Support）

支持度表示项集{X,Y}在总项集里出现的概率。公式为：

$$Support(X \rightarrow Y) = P(X,Y) / P(I) = P(XUY) / P(I) = num(XUY) / num(I)$$

其中，I表示总事务集。num()表示求事务集里特定项集出现的次数。

比如，num(I)表示总事务集的个数

num(XUY)表示含有{X,Y}的事务集的个数（个数也叫次数）。

2.置信度 （Confidence）

置信度表示在先决条件X发生的情况下，由关联规则"X→Y"推出Y的概率。即在含有X的项集中，含有Y的可能性，公式为：

$$Confidence(X \rightarrow Y) = P(Y|X) = P(X,Y) / P(X) = P(XUY) / P(X)$$

3.提升度（Lift）

提升度表示含有X的条件下，同时含有Y的概率，与不含X的条件下却含Y的概率之比。

$$Lift(X \rightarrow Y) = P(Y|X) / P(Y)$$

它表明提升的程度与置信度成正比。当 Y 的支持度不变时，提升度与置信度成正比。线性趋势的斜率是支持度（y）的倒数。

It shows that the degree of lift is proportional to the degree of confidence. When the support degree of Y remains unchanged, the lift degree is directly proportional to the confidence level. The slope of the linear trend is the reciprocal of support (y).

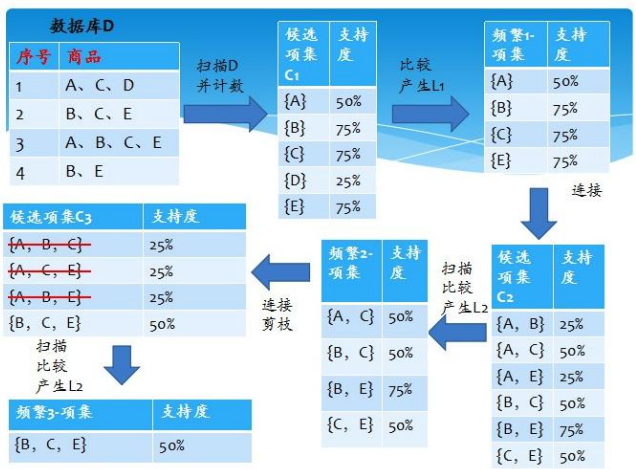
C. 解释关联算法的先验方法，关键步骤 ？

1、支持度
关联规则中 A->B 的支持度，指 AB 同时发生的概率
2、置信度
置信度 confidence=P(B|A)=P(AB)/P(A),指的是发生事件 A 的基础上发生事件 B 的概率
3、K 项集
事件 A 中包含 k 个元素，那么称这个事件 A 为 k 项集，并且事件 A 满足最小支持度阈值的事件称为频繁 k 项集
4、由频繁项集产生的强关联规则

D. Apriori 算法

Apriori 算法是一种最有影响力的挖掘关联规则的频繁项集的算法
它使用一种称作逐层搜索的迭代方法，k 项集用于探索 k+1 项集
Apriori algorithm is one of the most influential algorithms for mining **frequent itemsets** of **association rules**
It uses an **iterative method** called layer by layer search. K itemsets are used to explore (K + 1) itemsets

Apriori 算法过程分为三个步骤：
第一步，先要设置最小支持度阈值，然后查找出数据库中的所有频繁项集；
第二步，利用频繁项集生成新的候选集，并计算他们的支持度；
第三步，根据最小支持度阈值挑选出新的频繁集直到无法发现更多频繁项；
The first step is to set the minimum support threshold, and then find out all the frequent itemsets in the database;
In the second step, the frequent itemsets are used to generate new candidate sets and their support is calculated;
In the third step, new frequent sets are selected according to the minimum support threshold until no more frequent items can be found;



两大定律：
1、如果一个项集是频繁项集，那么它的所有子集也是频繁项集；
2、如果一个项集不是频繁项集，那么它的所有超集都不是频繁项集；
1. If an itemset is a frequent itemset, then all its subsets are also frequent itemsets;
2. If an itemset is not a frequent itemset, then all its supersets are not frequent itemsets;

Q6

A. 在训练完一个逻辑回归模型之后，描述你观察到什么?
可以发现，数据具有一定的线性相关性。数据集沿直线分布。
残差均匀分布在 x 轴周围。采用线性回归更为合适
It can be found that the data has a **certain linear correlation**. Data sets are distributed along a straight line.
The residuals are uniformly distributed around the x-axis. Linear regression is more suitable.

B. 描述逻辑回归如何被用于分类任务?
二分类时的函数为 sigmod 函数，多分类问题应使用 softmax 函数。
线性回归预测是连续的，而不是概率。
线性回归对不平衡数据非常敏感。
The function of binary classification is SIGMOD function, and softmax function should be used in multi classification
Linear regression prediction is continuous, not probabilistic.
Linear regression is very sensitive to unbalanced data.

C. 如果 b3 = -0.5，那么说明 X3 每增加一个单位，b3X3 就会减小一个单位，会使整个逻辑回归模型整体值 Y 下降，降低到 y=0 以下。关系式的斜率减小。
If B3 = - 0.5, it means that when X3 increases by one unit, b3x3 will decrease by one unit, which will reduce the overall value of Y in the whole logistic regression model to below y = 0. The slope of the relationship decreases.

D. 描述决策树用于选择信息量最大的属性的标准。
信息量最大的属性由信息增益（Information gain）标识，信息增益基于熵计算。
The attribute with the largest amount of information is identified by information gain, which is calculated based on **entropy**.

Q7

A. 描述将 Bayes 定理转换为朴素贝叶斯分类器的两种简化方法。
1. 使用条件独立性假设，每个属性是条件独立与其他属性的。
2. 忽略分母
1.Using conditional independence assumption, each attribute is conditionally independent from other attributes
2.to ignore the **denominator.**

B . 面对一堆充满不相关、无类别数据的分类任务时，Which classifier will you consider using?
(逻辑回归不适合分类问题，贝叶斯需要变量没有关联性，所以使用决策树)
Naive Bayes 需要假设数据变量是有条件独立的。Bayes 不适用于相关变量。分类器可以处理不同层次的 logistic 回归。我们记得决策树也可以处理分类变量，但是太多的层次会导致树的深度。
Naive Bayes needs to assume that data variables are conditionally independent. Bayes is not suitable for correlated variables. Different from logistic regression, naive Bayes classifier can deal with multi-level classification variables. We remember that decision trees can also handle categorical variables, but too many levels can lead to deep trees.

C. 是否评论，评论好，评论坏

首先，需要收集数据。然后对信息进行分类。如果数据之间没有相关性，可以使用 naivebayes。

如果数据之间存在相关性，可以采用决策树方法进行数据处理。

First, datasets needs to be collected. Then classify the information. If there is no correlation between the data, naive Bayes can be used. If there is correlation between data, the method of decision tree can be used for data processing.

决策树：如何防止过拟合

1. 提前停止
   - 限制决策树的高度，可以利用**交叉验证**的方法
   - 限制分类条件，如果下一次切分没有降低误差，则停止分支
   - 限制节点个数
2. 剪枝
   - 后剪枝（Post-Pruning）
     - 先构建完整的决策树，允许决策树过度拟合训练数据。
     - 然后对那些置信度不够的节点的子树用叶节点来替代
     - 该叶节点持有其子树的数据集中样本最多的类或者其概率分布。

Decision tree: how to prevent over fitting

1. Stop ahead of time
   - **Cross validation** can be used to limit the height of decision tree
   - If the next segmentation does not reduce the error, the branch will be stopped
   - Limit the number of nodes
2. Pruning
   - Post pruning
   - Firstly, a complete decision tree is constructed to allow the decision tree to over fit the training data.
   - Then, the subtree of nodes with insufficient confidence is replaced by leaf nodes
   - The leaf node holds the most sample class or its probability distribution in the data set of its subtree.

决策树的构造

决策树学习的算法通常是一个递归地选择最优特征

根据该特征对训练数据进行分割，使得各个子数据集有一个最好的分类的过程。

这一过程对应着对特征空间的划分，也对应着决策树的构建。

construction of decision tree

1. The algorithm of decision tree learning is usually a process of recursively selecting the optimal feature.

2. Devide the training data according to the feature, so that each sub data set has a best classification process.

3. This process corresponds to the division of feature space and the construction of decision tree.

1）开始：构建根节点，**将所有训练数据都放在根节点**，**选择一个最优特征**，按着这一特征将训练数据集**分割**成子集
2）如果这些子集**已经能够被基本正确分类，那么构建叶节点**，并将这些子集分到所对应的叶节点去
3）如果还有子集不能够被正确的分类，对这些子集重新选择新的最优特征，继续对其进行分割，构建相应的节点，如果递归进行，直至所有训练数据子集被基本正确的分类，或者**没有合适的特征为止**
4）每个子集都被分到叶节点上，即都有了明确的类，这样就生成了一颗决策树

1) Start: construct t**he root node**, put all the training data in the root node, select best feature, **divide the training data set into subsets** according to this feature
2) If these subsets can be basically correctly classified, then **build leaf nodes** and divide these subsets into corresponding leaf nodes.
3) If there are subsets that can not be correctly classified, then the new optimal features are selected for these subsets, and then the corresponding nodes are constructed. If recursion is carried out, until all training data subsets are basically correctly classified or there are no **suitable features**.
4) Each subsets is divided into leaf nodes, that is to say, there are definite classes, thus a decision tree is generated.

贝叶斯

Q8

A. 文本分析的步骤
1.解析：获取非结构化文本并转换成结构化，数据预处理，转化为词向量
2.搜索和**检索**：识别包含搜索项的语料库中的文档
3.文本挖掘：利用聚类和分类算法发现有意义的结果，利用 k-means 将文件分组，用朴素贝叶斯分类器进行情感分析
1. Parse: get unstructured text and convert it into structured text
2. Search and **retrieval**: identify documents in corpus containing search terms
3. Text mining: use clustering and classification algorithm to find meaningful results, use k-means to group files, and naive Bayes classifier to analyze emotion

B. 词袋描述--》文档中的 N 个文字转化为 N 高维向量　Bag-of-words

Bag-of-words 模型是常用的信息检索方法来表示文档。
它将文档成为高维向量，表示文档中各种单词的存在/缺失/频率。
Bag-of-words model is a common **information retrieval method** to represent documents。
It makes the document into a **high-dimensional vector**, indicating the presence / absence / frequency of various words in the document。

比如说：在一个巨大的文档集合 D，里面一共有 M 个文档，而文档里面的所有单词提取出来后，一起构成一个包含 N 个单词的词典，利用 Bag-of-words 模型，每个文档都可以被表示成为一个 N 维向量。
变为 N 维向量之后，很多问题就变得非常好解了，计算机非常擅长于处理数值向量，我们可以通过余弦来求两个文档之间的相似度，也可以将这个向量作为特征向量送入分类器进行主题分类等一系列功能中去

C. **TF-IDF 原理和应用**

TF-IDF(Term Frequency-Inverse Document Frequency, 词频-逆文件频率).

词频（TF）　　某个词在文章中的出现次数，这个数字通常会被归一化
文件中术语的流行程度　Popularity of the term in the document

$$TF_w = \frac{\text{在某一类中词条}w\text{出现的次数}}{\text{该类中所有的词条数目}}$$

逆向文件频率（IDF）
语义库中术语的稀缺性　The rareness of the term in semantic database

$\text{IDF} = \log(\frac{\text{语义库中文档总数}}{\text{包含词条 w 的文档数+1}})$，**分母+1 是因为避免分母为 0**
The denominator + 1 is to avoid a denominator of 0

TF-IDF = 词频（TF）* 逆文档频率（IDF）

TF-IDF 是一种统计方法，用以评估一个词对于一个文件集，或着语料库中的一份文件的重要程度。
TF-IDF is a statistical method to evaluate the importance of a word to a document set or a document in a corpus.

字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。
一个词语在一篇文章中出现次数越多，同时在所有文档中出现次数越少，越能够代表该文章.
The importance of a word increases in proportion to the number of times it appears in the document, but at the same time it decreases inversely with the frequency of its appearance in the corpus.
The more times a word appears in an article, the less it appears in all documents, the more likely it is to represent the article

TF-IDF 的特点：
•TF-IDF 完全基于所有获取的文档
•一旦获取的文档发生更改，可以轻松地更新 TFIDF
•TF-IDF 是文本分析中广泛使用的一种度量方法

- TF-IDF is based entirely on all the fetched documents
- TF-IDF can be easily updated once the fetched documents change
- TF-IDF is a measure widely used in text analysis

很多停止语词语可以直接去掉，因为不起作用：例如 the , a, of , and to，语义理解 - semantic understanding

一个词语在一篇文章中出现次数越多，同时在所有文档中出现次数越少，越能够代表该文章.

词频（term frequency, TF）指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化(一般是词频除以文章总词数)，以防止它偏向长的文件。

逆向文件频率（inverse document frequency, IDF) IDF 的主要思想是：
如果包含词条 t 的文档越少, IDF 越大，则说明词条具有很好的类别区分能力。
某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。

	包含该词的文档数（亿）	IDF	TF-IDF
中国	62.3	0.603	0.0121
蜜蜂	0.484	2.713	0.0543
养殖	0.973	2.410	0.0482

从上表可见，"蜜蜂"的TF-IDF值最高，"养殖"其次，"中国"最低。（如果还计算"的"字的TF-IDF，那将是一个极其接近0的值。）所以，如果只选择一个词，"蜜蜂"就是这篇文章的关键词。

D. 按主题对文档进行分类 Categorizing Documents by Topics

注：topic：文章的题目，theme: 文章的中心思想
按主题对文档进行分类，可以使用上文提到过的 TF-IDF 方法：

为什么用 TF-IDF 来进行主题文档分类 ？
- 提供相对较少的描述长度
- 很少显示**文档间和文档内的统计结构**
- Provides relatively small amount of reduction in description length
- reveals little inter-document or intra-document statistical structure

什么是主题 ？ Topic

1. 主题就是经常出现在一起并共享同一个主题的一组词
每个词在这个主题中都有权重
• A topic: **a cluster of** words that frequently occur together and share the same theme
Each word has a weight inside this topic

2. 主题被正式定义为固定词汇表上的分布
同一词汇表中不同的主题有不同的分布
• A topic is formally **defined as a distribution over a fixed vocabulary of words**
Different topics have different distributions over the same vocabulary

3. 一个主题可以看作是一组具有相关含义的单词
词汇表中的一个词可以位于多个主题中，具有不同的权重
• A topic can be viewed as a cluster of words with related meanings
A word from the vocabulary can reside in multiple topics with different weights

主题模型的概念和优点：

- •主题建模为文档提供简短描述
- •帮助组织、搜索、理解和总结文本
- •主题模型是统计模型
  - 检查一组文档中的单词
  - 确定课文的主题
  - 探索主题之间的关联

Concept and advantages of topic model：

- Topic modelling provides **short descriptions** for documents
- Helps to organize, search, understand, and summarize text
- Topic models are statistical models that
  - Examine words from a set of documents
  - Determine the themes over the text
  - Discover how themes are associated

E. Determining Sentiments

情感分析使用统计和自然语言处理来从文本中挖掘观点来识别并抽取主观信息。 分类器仅基于对其进行训练的数据集来确定情感:

Sentiment analysis uses statistics and NLP to mine opinions to identify and exact **subjective information** from texts

- 词义随着领域不同而改变。
- 因此模型无法直接应用于其他领域。 绝对的情感水平无法提供信息，应该建立一条基线，然后将其与观察

值比较。

The meaning of a word changes with the field.

Therefore, the model can not be directly applied to other fields. Absolute emotional levels do not provide information, and a baseline should be established and then compared with the observed values.

Q9. 图像分析

•图像分析概述
指视觉数据的表示、处理和建模，以获得有用的见解
• Overview of Image Analysis
Refers to the representation, processing, and modelling of visual data to derive useful insights

表示图片：
传统表示是手动选择的特征，全局特征（如强度，颜色，纹理，形状和结构等）。
Represents the picture:
Traditional representations are manually selected features, global features (such as strength, color, texture, shape and structure).

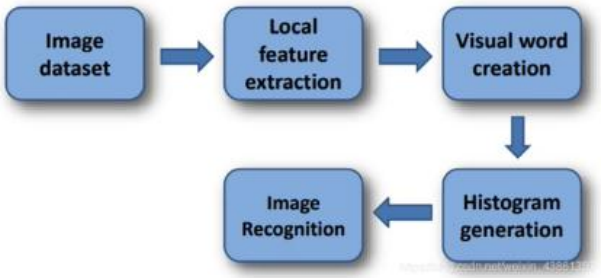
后来的特征提取方法是 SIFT, HOG 等方式来表示图片，其对角度，比例，明度等都是具有不变性。
SIFT ： scale invariant feature transform 尺度不变特征变换
HOG：histogram of oriented gradients 定向梯度直方图
这两种方法都是基于图像中梯度的方向直方图的特征提取方法。

最新的特征提取方法是基于深度学习的。

Later feature extraction methods include SIFT, HOG and other methods to represent the image, which are invariant to Angle, proportion and brightness, etc.

Both of these methods are feature extraction methods based on orientation histogram of gradient in image.

The latest feature extraction method is based on deep learning.



- **深度神经网络识别图片 Deep Convolutional Neural Networks**

卷积神经网络 CNN 是一种前馈网络，可以从图像中提取拓扑属性。像几乎所有其他神经网络一样，它也使用反向传播算法进行训练。

卷积神经网络旨在通过使用最少的预处理（**minimal preprocessing**）直接从像素图像中识别视觉模式（visual patterns）。他们可以识别具有极大可变性（**extreme variability**）的模式（例如手写字符）。

Convolutional neural network CNN is a **feed forward network**, which can extract **topological properties** from images. Like almost all other neural networks, it also uses **back-propagation** algorithm for training.

Convolution neural network aims to recognize visual patterns directly from pixel images by using minimal preprocessing. They can recognize patterns with extreme variability, such as handwritten characters.

Q10. 时间序列分析 Time Series Analysis

时间序列：等间距随时间的有序序列
An ordered sequence of equally spaced values over time

通过对一个区域进行一个时间段内的连续观测，提取图像有关特征，并分析其变化过程与预测发展规模
Through continuous observation of a region for a period of time, relevant features of the image are extracted, and analyze the change process to predict the development scale

时间序列分析方法：
1.条件数据和选择模型
- 确定并说明时间序列中的任何趋势或季节性
- 检查剩余时间序列并确定合适的模型

2.估算模型参数
3.评估模型，如果需要，返回步骤 1

1.Condition data and select a model
- Identify and account for any trends or seasonality in the time series
- Examine the remaining time series and determine a suitable model

2.Estimate the model parameters
3.Assess the model and return to Step 1, if needed

时间序列可以包括 – 趋势、季节性、周期性和随机性
A time series can consists of – Trend, Seasonality, Cyclic, and Random

•趋势
– 时间序列中的长期移动 – 值随时间增加或减少
•季节性
– 固定的、周期性的随时间变化的波动 – 通常与日历有关
•周期性
– 周期性的，但不是固定的随时间变化的波动 – 比如说，经济的繁荣-萧条周期
•随机
– 完成上述三个部分后，剩下的是什么
– 噪声+待建模的底层结构

ARIMA 模型 =» 用于时间序列预测分析
应适用于静止时间序列 =» Shall be applied to stationary time series

常用的时间序列模型有四种：自回归模型 AR(p)、移动平均模型 MA(q)、自回归移动平均模型 ARMA(p,q)、自回归差分移动平均模型 ARIMA(p,d,q)，可以说前三种都是 ARIMA(p,d,q)模型的特殊形式

ARIMA Model

•自回归（AR）模型
– 对于平稳时间序列，AR（p）表示为

- Autoregressive (AR) Models
  – For a stationary time series, AR(p) is expressed as

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$

where δ is a constant for a nonzero-centered time series:
ϕ_j is a constant for j = 1, 2, . . ., p
y_{t-j} is the value of the time series at time $t - j$
$\phi_p \neq 0$
$\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ for all t

点yt是先前p值的线性组合
A point y_t is a linear combination of the prior p values

- Moving Average (MA) models
  – For a time series, y_t, centred at zero, a MA(q) is expressed as

•移动平均（MA）模型
– 对于以零为中心的时间序列yt，MA（q）表示为

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

where θ_k is a constant for k = 1, 2, ..., q
$\theta_q \neq 0$
$\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ for all t

- AR(p) and MA(q) are often combined into one model for time series, resulting in ARMA(p,q).

•AR（p）和MA（q）通常被组合成一个时间序列模型，从而产生ARMA（p，q）。

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

where δ is a constant for a nonzero-centered time series
ϕ_j is a constant for j = 1, 2, ..., p
$\phi_p \neq 0$
θ_k is a constant for k = 1, 2, ..., q
$\theta_q \neq 0$
$\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ for all t

- ARIMA model
  – Autoregressive Integrated Moving Average
  – Differencing is included in ARMA model
- ARIMA(p,d,q)
  – ARMA(p,q) mode is applied after applying differencing d times

•ARIMA模型
– 自回归综合移动平均值
差分模型包含在ARMA中
•ARIMA（p、d、q）
– 应用差分d次后应用ARMA（p，q）模式

ARIMA 建立步骤：

步骤 1：获取被观测系统时间序列数据；
步骤 2：对数据绘图，观测是否为平稳时间序列（一般都不平稳）；对于非平稳时间序列要先进行 d 阶差分运算，化为平稳时间序列；
画出 ACF 图，由 ACF 图可以看出，自相关系数长期大于 0，说明序列具有很强的长期相关性，所以趋势并不平稳，因此要做差分运算，依次测试。
步骤 4：拟合 ARIMA 模型（0,1,1）
步骤 5：预测

Step 1: obtain the time series data of the observed system;
Step 2: plot the data, whether the observation is a **stationary time series** (generally unstable); for the non-stationary time series, the d-order difference operation should be carried out to **convert it into** a stationary time series;
Draw the ACF diagram. From the ACF diagram, we can see that the long-term autocorrelation coefficient is greater than 0, indicating that the sequence has a strong long-term correlation, so the trend is not stable, **so we need to do differencing operation and test in turn.**
Step 4: fitting ARIMA model (0,1,1)
Step 5: Forecast

- Linear Regression

线性回归是一种用来对若干输入变量与一个连续结果变量之间关系建模的分析技术。
An analytical technique used to model the relationship between several input variables and a continuous outcome variable
线性回归模型在普通最小二乘法（OLS）的基础上进行了额外的假设
Linear regression model makes additional assumptions on top of the **Ordinary Least Squares** (OLS)

- Logistic Regression

当结果变量是分类型的，逻辑回归用来基于输入变量预测结果的可能性。
When the outcome variable is categorical in nature, logistic regression can be used to predict the probability of an outcome based on the input variables

最大似然估计（MLE）常用于对模型参数进行估计，即找出观测给定数据集的机会最大的参数值
Maximum Likelihood Estimation (MLE) is often used to estimate the model parameters, which finds the parameter values maximizing the chances of observing the given dataset