1. 单选题 （每题2分，共15题）
2. 判断题 （每题2分，共5题）

1.当皮尔逊相关系数为0时，两个变量独立。答案用白色打出在两个括号之间（有可能是非线性相关，而非两个独立变量）

2.对于情感词抽取，基于统计的方法准确率高，而基于规则的方法召回率高。（错，统计召回率高，规则准确率高）

3.在寻找相似文档时在shingling这步k值得选取十分重要，其中对于段文档k值应该选取3，长文档k值选取10（错，段文档k值选5）

4.幂律分布描述的是一种少数元素占有多数资源的规律。（对 ）

5.图的直径是节点对间距的平均值（错，最大取值）

6.真实世界的图规律：随着节点数增加，有效直径随之增加。（错，有效直径下降）

1. 简答题 （每题6分，共3题）

给出并简要解释xxx模型中的损失函数。

1. 请简要解释可拓展算法，并说明什么是线性可拓展和超可拓展。
2. 请说明参数更新的三种方案和其优缺点。
3. 请用伪代码简要说明MapReduce的过程。
4. 请分别解释，熵，条件熵，信息增益的含义并给出其表达式。
5. 比较说一下 Random Forest 与 GBDT的共同点与区别？
6. 简要描述近似求解最小二乘法的过程。
7. 请写出传统相关性分析的三个相关系数的计算方法（肯德尔系数、皮尔森系数、斯皮尔曼系数）

8.简要描述隐性语义索引（LSI）的过程

9.请给出排序学习模型（c&w）的损失函数和其参数的含义

10.上下文预测模型Word2Vec的CBOW和SG模型的目标函数分别时什么

11.全局上下文模型GloVe对单词的软约束和目标函数是什么

12.TF-IDF模型的TF和IDF怎么计算

13.请写出文本匹配的评价方法中nDCG的指标定义

14.请写出张量分解的目标函数并简要解释其含义

15请写出基于分布式表达的推理：TransE的损失函数并简要解释其含义

1. 证明题/应用题（前三题每题10分，最后一道题12分）

xxx模型是被广泛采用的实体抽取模型之一，请简要绘制该模型的结构示意图，并给出其

1. 请说明寻找相似文档的过程，并计算在b=20，r=5，s=0.8时两个文档相似的概率。
2. 证明为什么在SVD中U中第一列时矩阵A的行坐标在第一个右奇异向量V1上投影的坐标。
3. CMS模型是十分重要的概率数据结构，请简要绘制该模型的结构示意图。
4. FRAUDAR是一种线性启发式检测密集块的算法,请简要说明该算法的过程，并证明g（x）具有下界：g（x）>0.5g(opt)。
5. 请简要绘制PageRank中的死胡同和蜘蛛网模型，并简要描述解决办法。
6. 请简要绘制LSTM模型的结构图，并说明遗忘门、输入门和输出门的作用及其门控向量的公式。
7. 请简要说明PageRank完全算法的流程。