北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

文本聚类汇报

汇报人：欧林堉 韩文轩 钱驰骋 王博为 胡玉麟

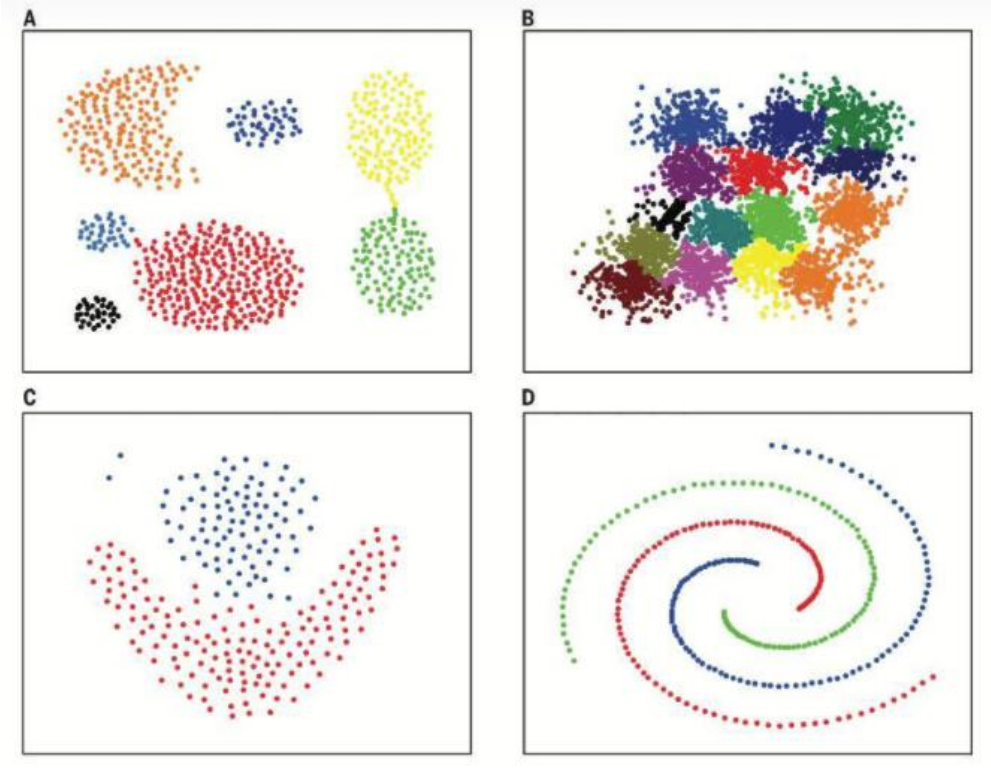
导　师：张华平

时　间：2022/4/17

目录 | CONTENTS

1 文本聚类概述

聚类：根据数据的不同特征，将其划分为不同的数据类。



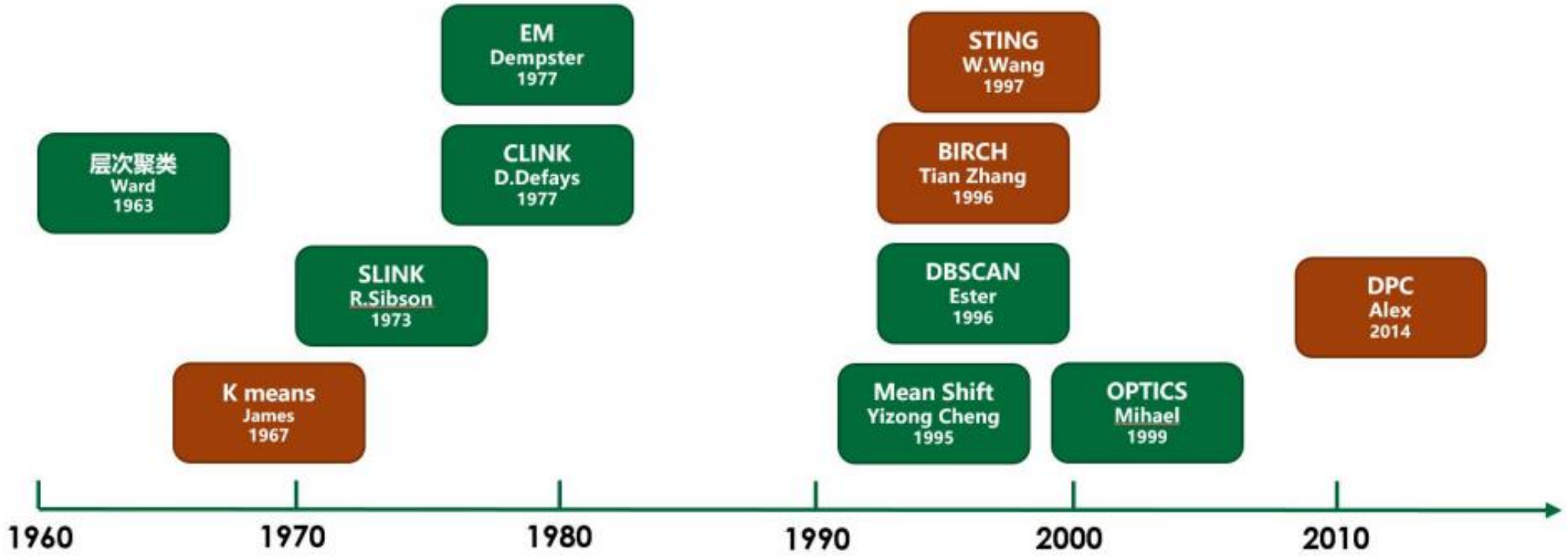
概述——基本概念

文本聚类：把相似度高的文本聚到一类，相似度低的文本分到不同的类。

例如：同为小说的《复活》和《战争与和平》相似度应该较高。而《工科数学分析》和《红与黑》的相似度应该较低。

聚类——聚类的发展

聚类算法是最早被用于模式识别及数据挖掘任务的方法之一，并且被用来研究各种应用中的大数据库，因此如今用于大数据的聚类算法受到越来越多的关注。



文本聚类与文本分类

文本聚类

文本聚类是一种无监督学习，数据不带标签。将相似的文本分为同一“簇”。文本聚类比较适合用于大数据中热点话题或事件的发现。

文本分类

文本分类是一种监督学习，数据带有标签，应用场景评论情感分析，新闻极性分析，新闻分类等等。

概述——应用场景

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。

文章推荐

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

热点追踪

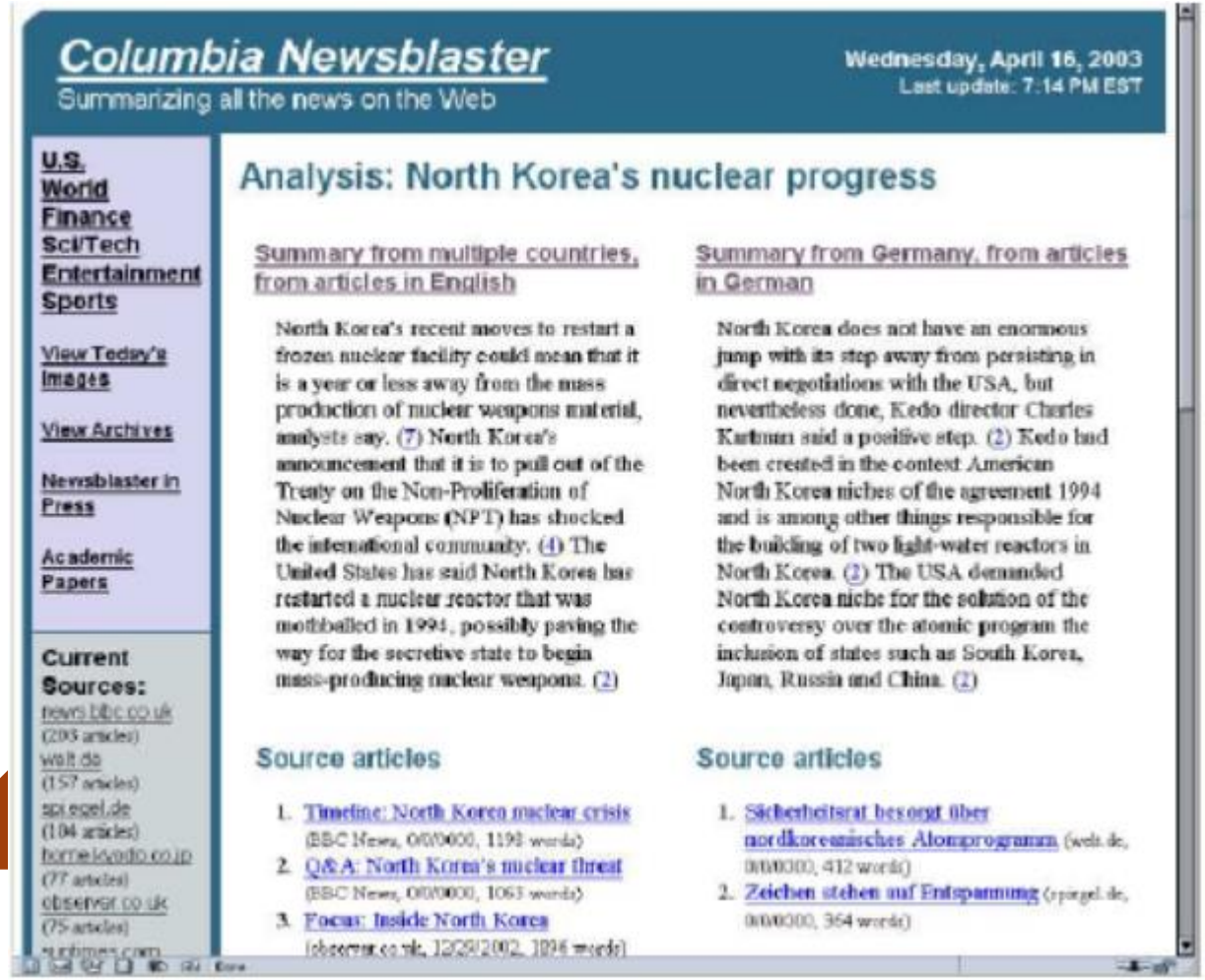
与文本分类不同，文本聚类不需要知道每个类别是什么，更适合网络热点追踪。

其他

搜索引擎联想功能，相关资源推荐推荐，改善文本分类的结果。

概述——应用场景

多文档自动文摘：Newsblaster：



概述——应用场景

搜索引擎返回结果：

概述——应用场景

数字图书馆：

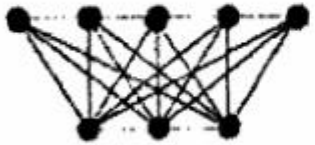


图 4.1.1 一维 SOM 结构图

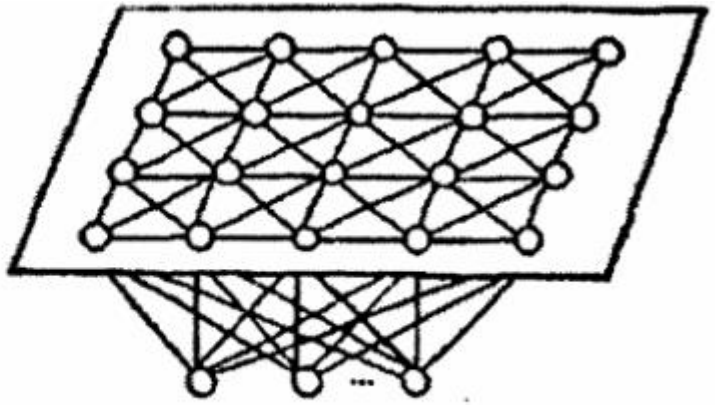


图 4.1.2 二维 SOM 结构图

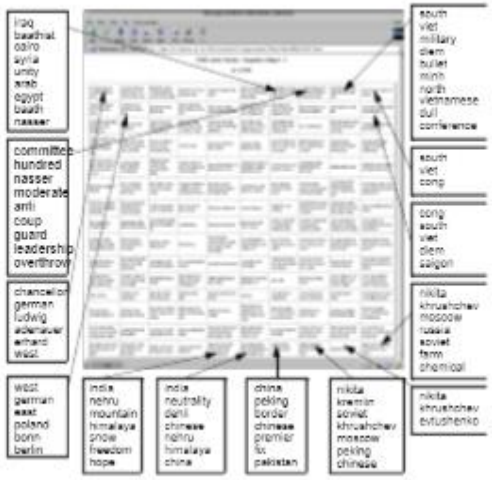


Figure 1: Labeled 10 15 SOM integrating 6 maps

Figure 2: Visualizing metadata of documents

概述——应用场景

文档集合自动整理：

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

文本预处理

文本预处理步骤将对文本进行分词和去停用词等处理。

文本表示

从文本中提取出特征，通过这些特征来表示相应的文本。

文本聚类

将上一步得到的用特征表示的文本送入聚类算法进行聚类，得到最终的文本聚类结果。

2 文本预处理与文本表示

文本表示的两种方法

文本的离散表示

one-hot，N-gram，TF-IDF存在着数据稀疏、向量维度过高、字词之间的关系无法度量的问题，适用于浅层的机器学习模型。

文本的分布式表示

根据上下文表示词语，是稠密、低维、连续的向量。

文本的离散表示（传统方法）

One-Hot

在语料库中,对每个词建立一个索引表示。

词袋表示

在 One-Hot 的基础上,用文档中各词的频数表示该文档。

N-gram

与词袋表示类似,将相邻 N 个单词编辑索引。

TF-IDF

在词袋表示的基础上对词进行 TF-IDF 值加权表示,TF-IDF 值与一个词在当前文档中出现频数成正比,与该词在整个语料库中出现频数成反比。

存在的问题：

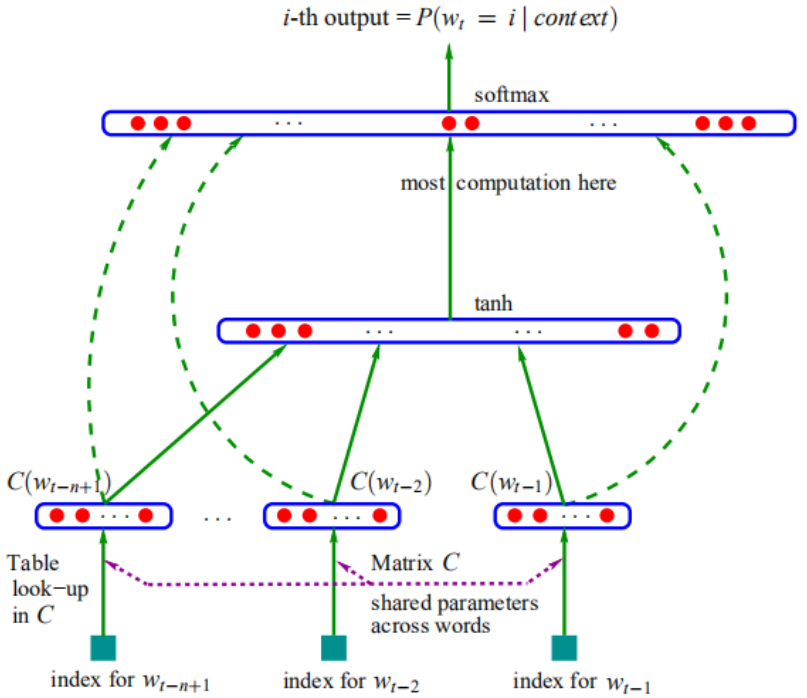
维度灾难、向量稀疏、不能捕捉长距离信息、不能表示文本潜在的语法与语义信息。

解决方案：

将高维向量映射为更加低维、稠密的连续向量的分布式表示方法。

NNLM

NNLM首次使用神经网络，利用前面 N 个词来预测当前词的概率模型，并将其预测概率最大似然化。
改进版有RNNLM、LSTM-RNNLM。

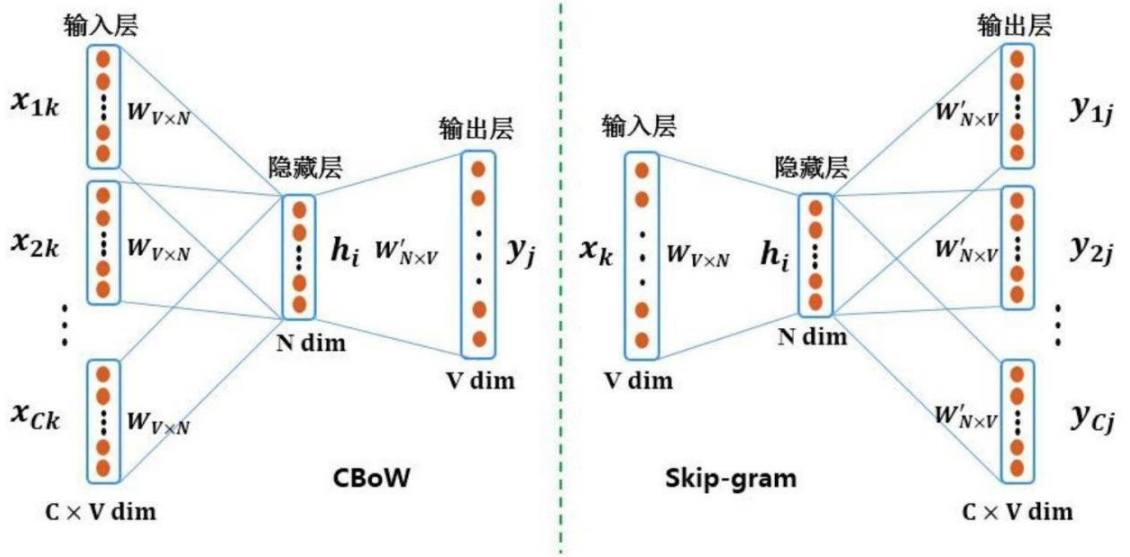


Word2vec

包含 CBOW 和 Skip-gram 两个模型,在CBOW 中,要求利用上下文词来预测中间目标词;在 Skip-gram 中,要求利用当前词来预测上下文词。

优点:
解决了处理变长序列的问题,并提出了两种模型训练提速方法,训练速度快。

不足:
无法解决一词多义问题,对分词结果敏感。



文本的分布式表示——继续进步

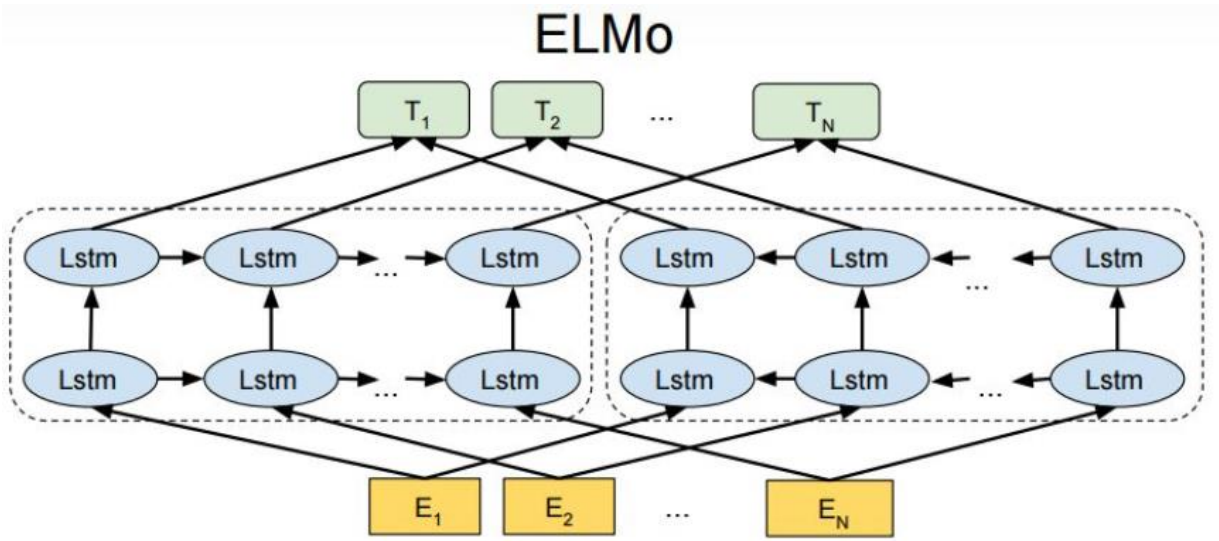
NNML和 Word2Vec属于静态词向量表示，无法解决一词多义问题,对分词结果敏感。

ELMo

在基础语言模型的词向量表示的基础上,采用双层双向 LSTM,分别从正反两方向对词汇进行编码。

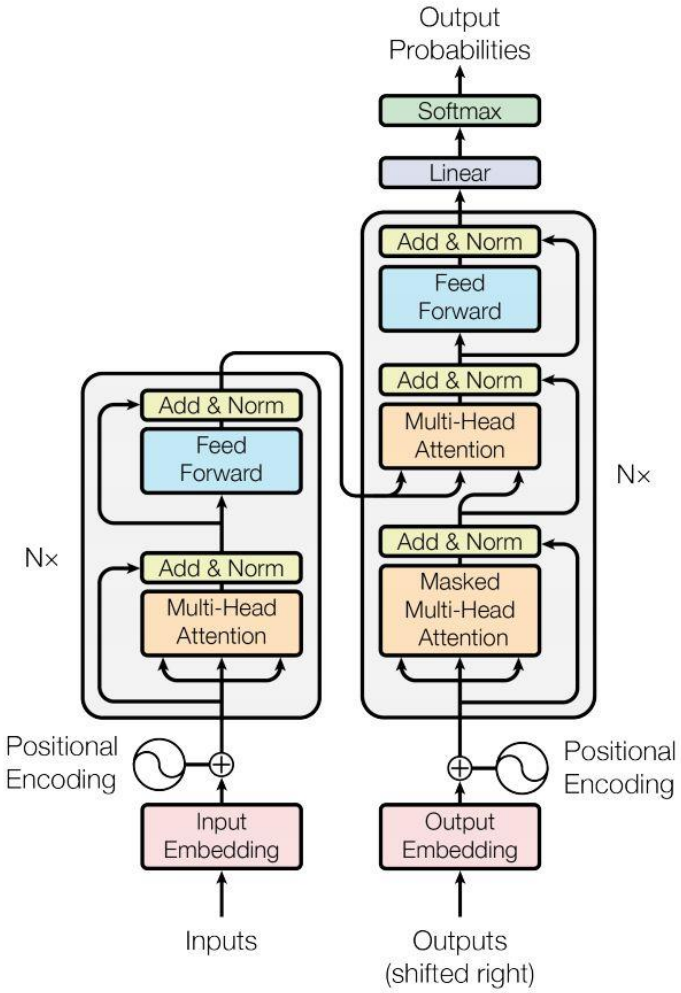
不足:

捕获长期依赖能力仍然有限。



transformer

2017年，Ashish Vaswani等人发表了《Attention is all you need》，推出了一个超越RNN的神经网络结构,即Transformer。
捕捉长距离文本信息的能力大大增强。
而且可以并行计算。



GPT

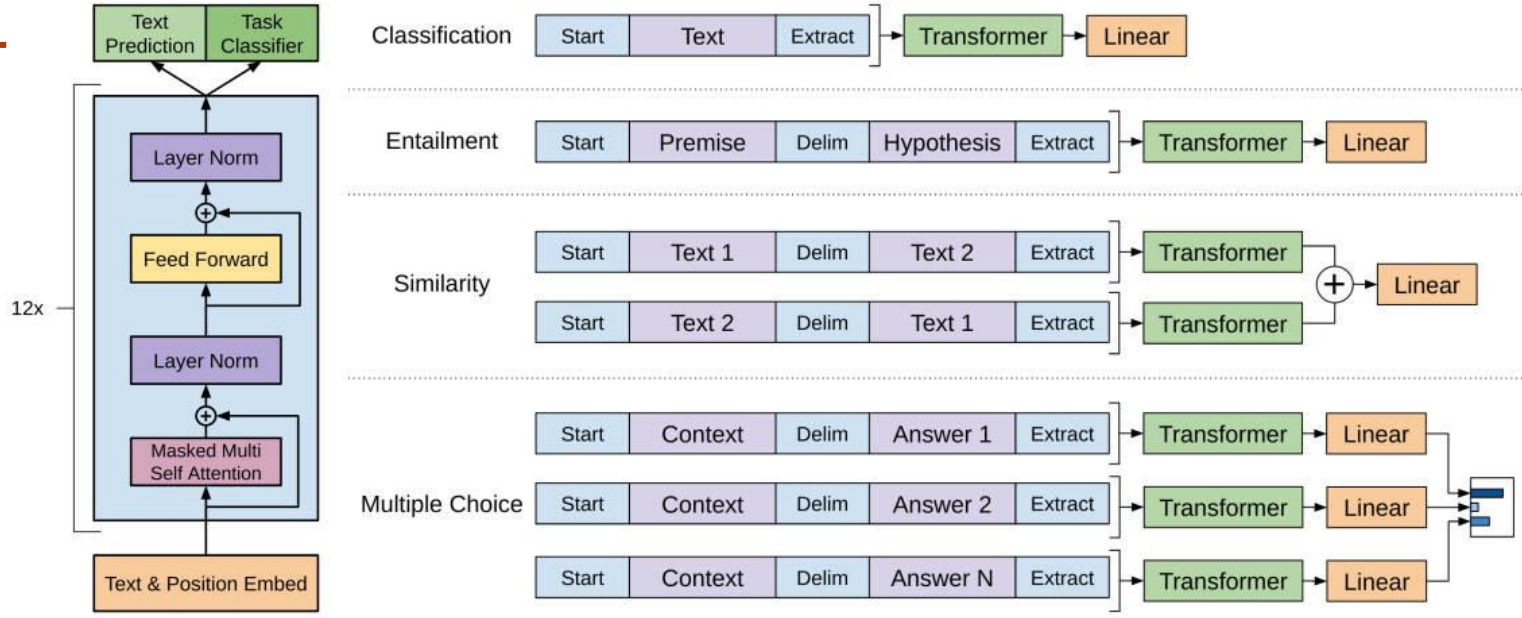
不同于 ELMo,GPT 利用 Transformer 结构进行单项语言模型的训练。

优点：

能够捕捉更长距离文本信息,语义关联效果更好,通过并行化实现训练提速,擅长处理文本生成任务。

不足：

只使用单向的语言表征信息,无法获取双向上下文信息表征。



Bert

将双向 Transformer 应用于语言模型,在双向上深度融合上下文特征。

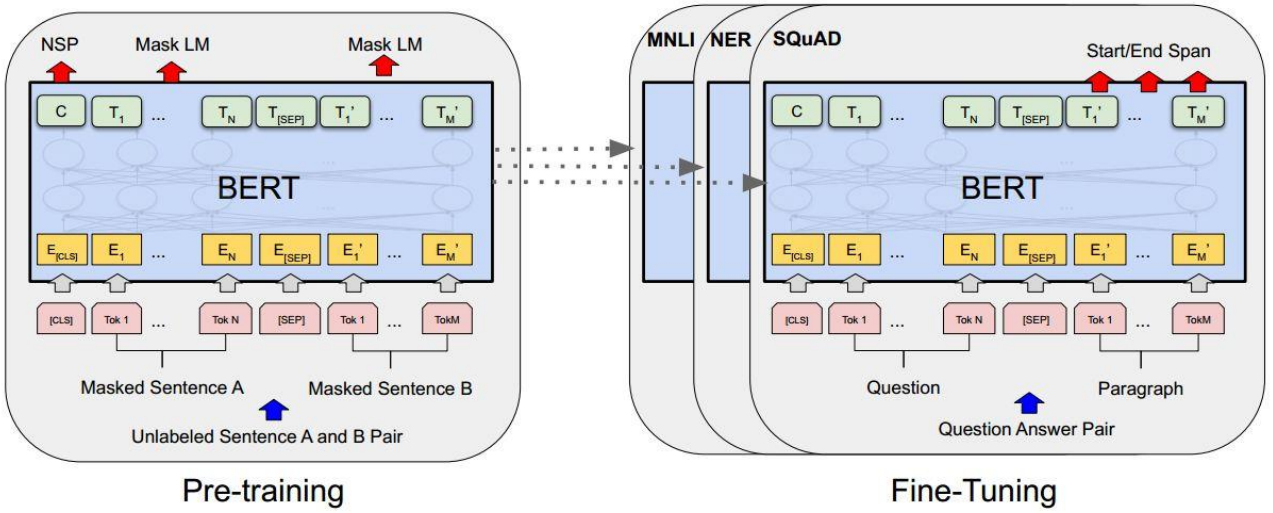
优点:

更好融合上下文特征信息,能够捕捉词语和句子级别的表示。

不足:

训练时使用的特殊标记“[MASK]”，导致在下游任务中训练与微调之间不匹配。

改进版有XLNet、RoBERTa、ALBERT、ERNIE。

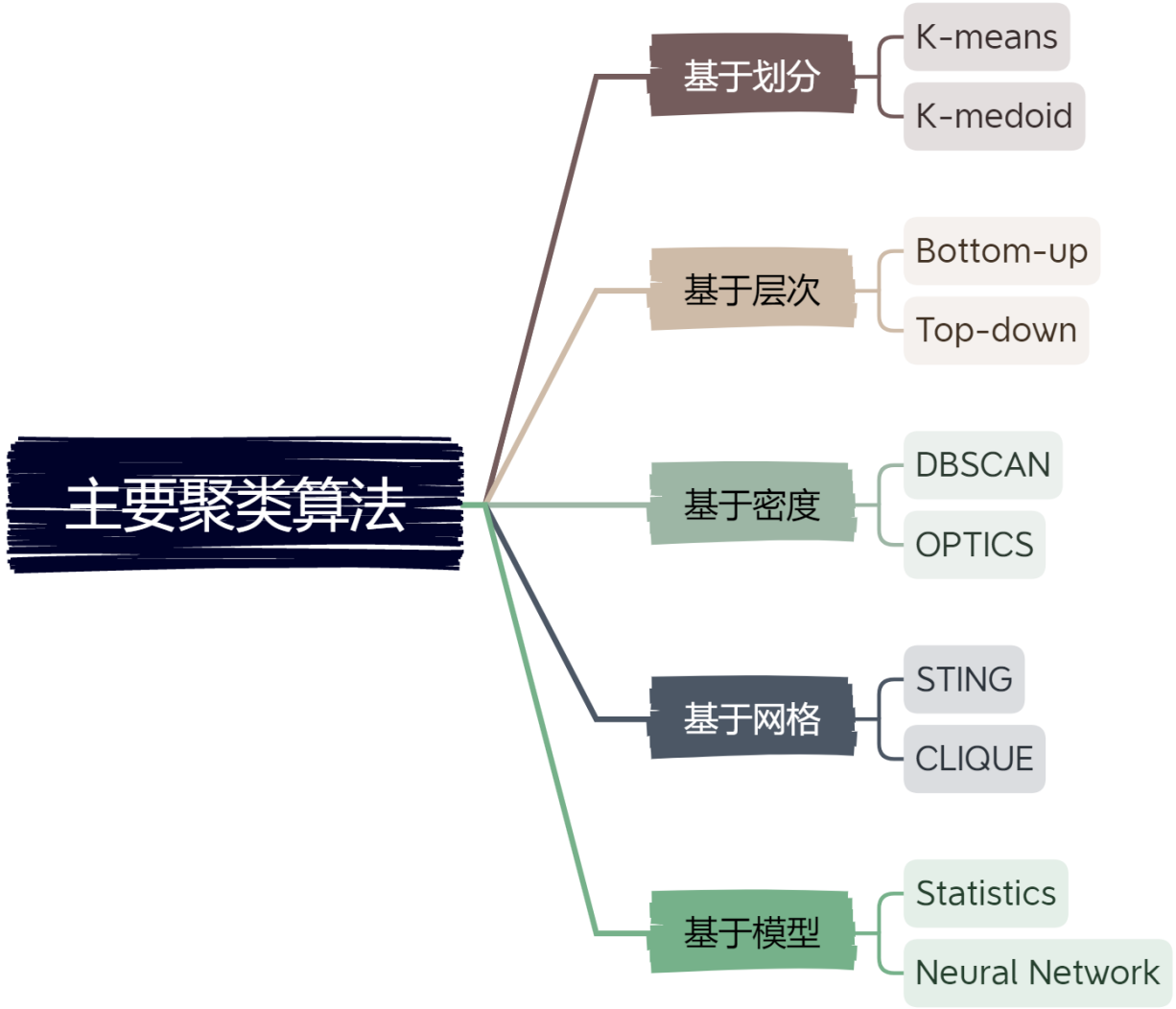


文本表示——总结

NNML和 Word2Vec属于静态词向量表示，无法解决一词多义问题,对分词结果敏感。

属于动态词向量的表示方法 ELMo、GPT和 BERT是在基础语言模型训练得到词向量的基础上,再在实际应用场景中对其进行动态调整,解决了静态词向量表示中的一词多义问题。

3 文本聚类


主要聚类算法
基于划分
K-means
K-medoid
基于层次
Bottom-up
Top-down
基于密度
DBSCAN
OPTICS
基于网格
STING
CLIQUE
基于模型
Statistics
Neural Network

距离！

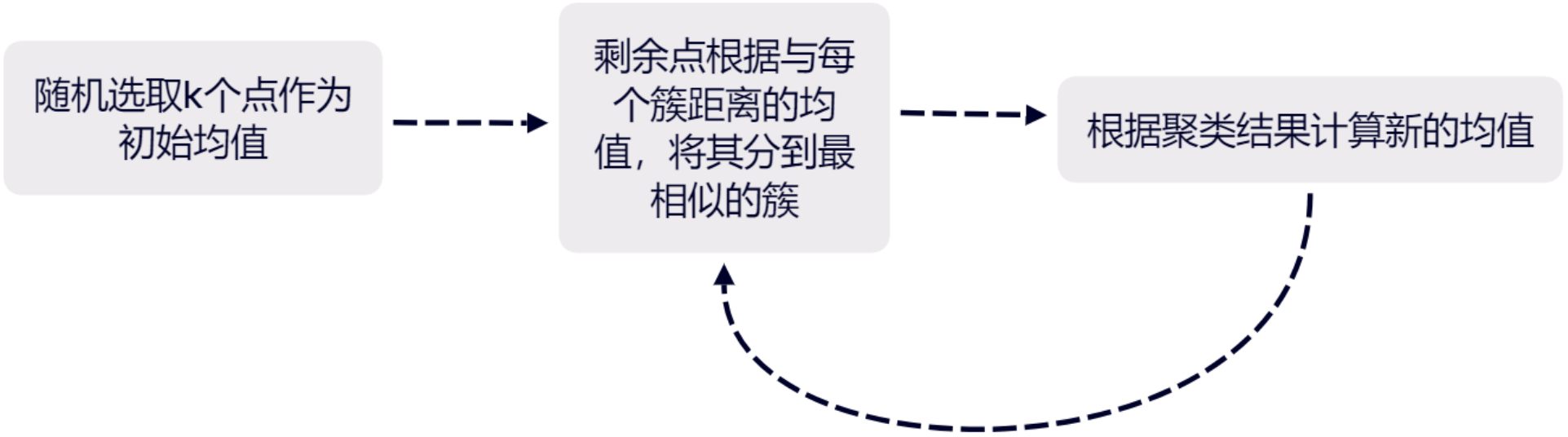
欧拉距离 $$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

曼哈顿距离 $$d = \sum_{i=1}^{n}|x_i - y_i|$$

闵氏距离 $$d = \sqrt[p]{\sum_{i=1}^{n}(x_i - y_i)^p}$$

Jaccard 相似系数 $$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

以簇数目k为输入参数，把n个对象划分成k个簇，使得簇内的相似度高，而簇间的相似度低



迭代直到收敛

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

初始化

1. 从样本集X中随机选择一个样本点作为第1个聚类中心；
2. 计算其它样本点x到最近的聚类中心的距离 d(x)；
3. 以概率 $\frac{d^2(x)}{\sum d^2(x)}$ 选择一个新样本点，加入聚类中心点集合中

缺点：受数据中的噪声和孤立点影响大！！

K-mediod

剩余点根据与每个簇距离的均值，将其分到最相似的簇

改进

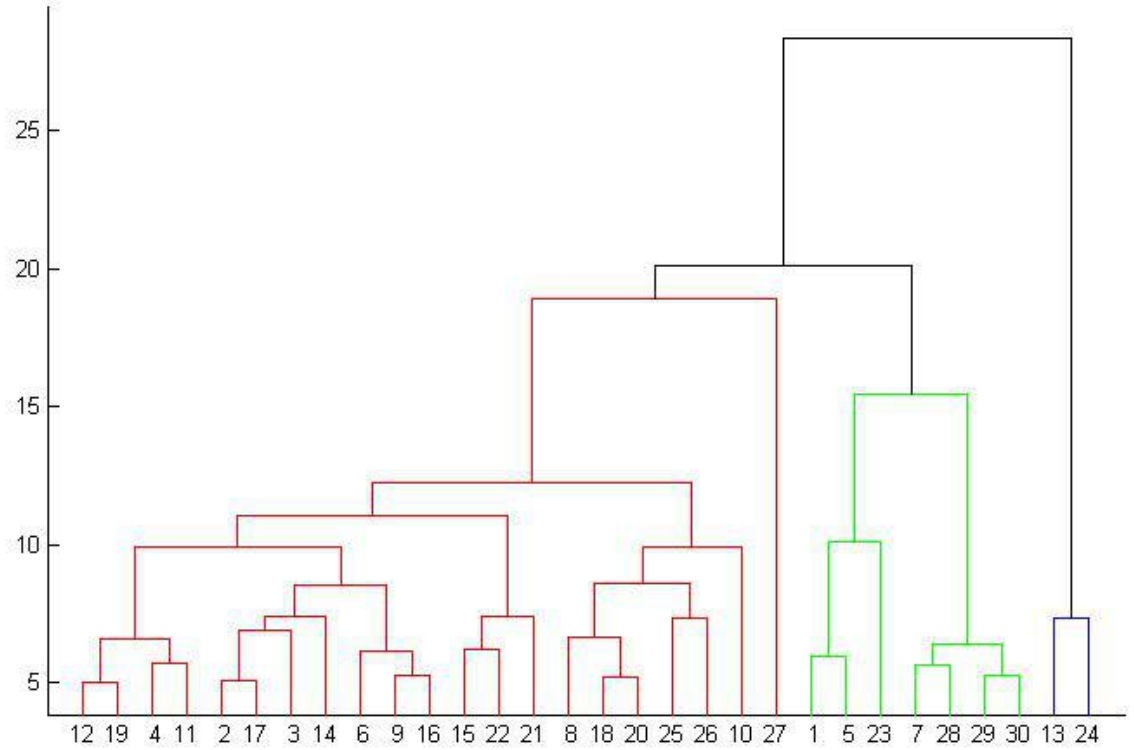
对每个类中除类中心的点外的其他样本点计算到其他所有点的距离和的最小值，将该最小值点作为新的聚类

但时间复杂度提高，不适合**大数据**!

AGNES

计算任意两个簇的距离，并找到最近的两个簇，合并。

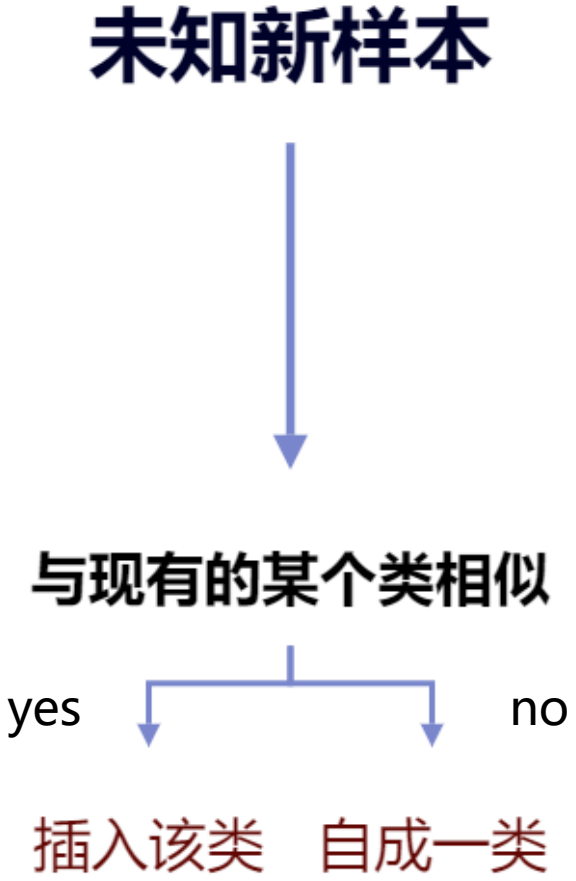
算法性能：
（1）简单，但有合并点选择困难的情况。
（2）一旦一组对象被合并，不能撤销
（3）算法复杂度为O(n^2)



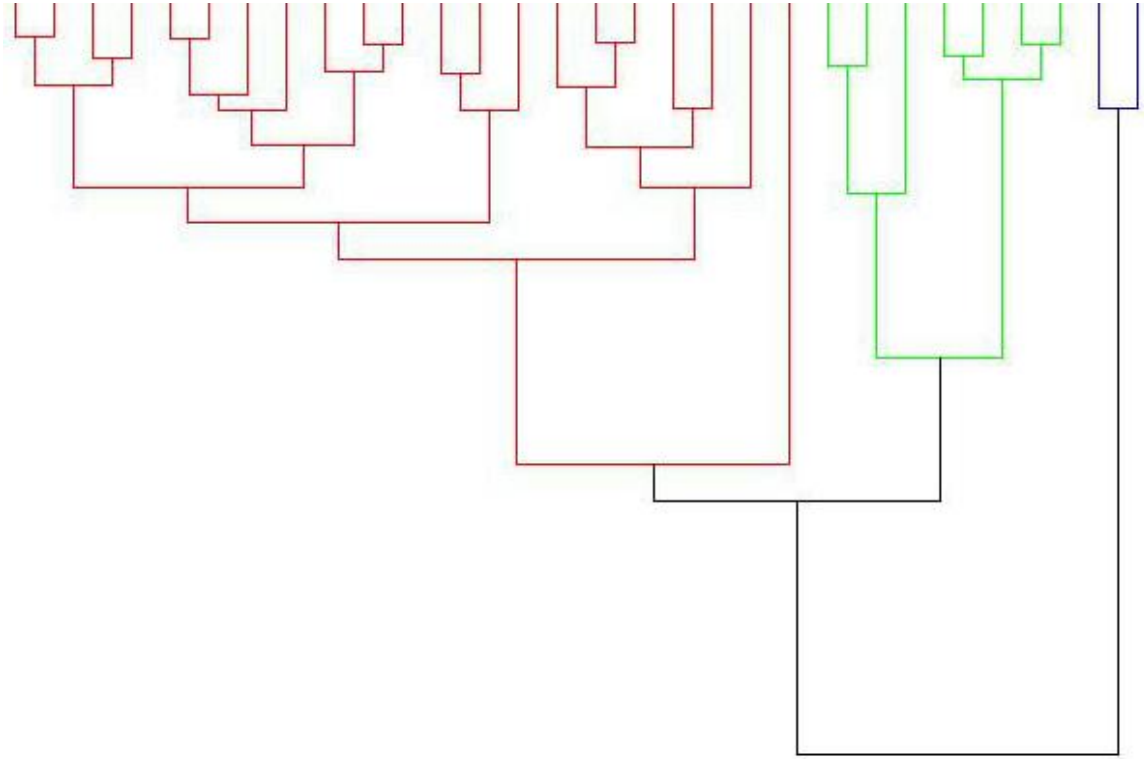
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Single-Pass

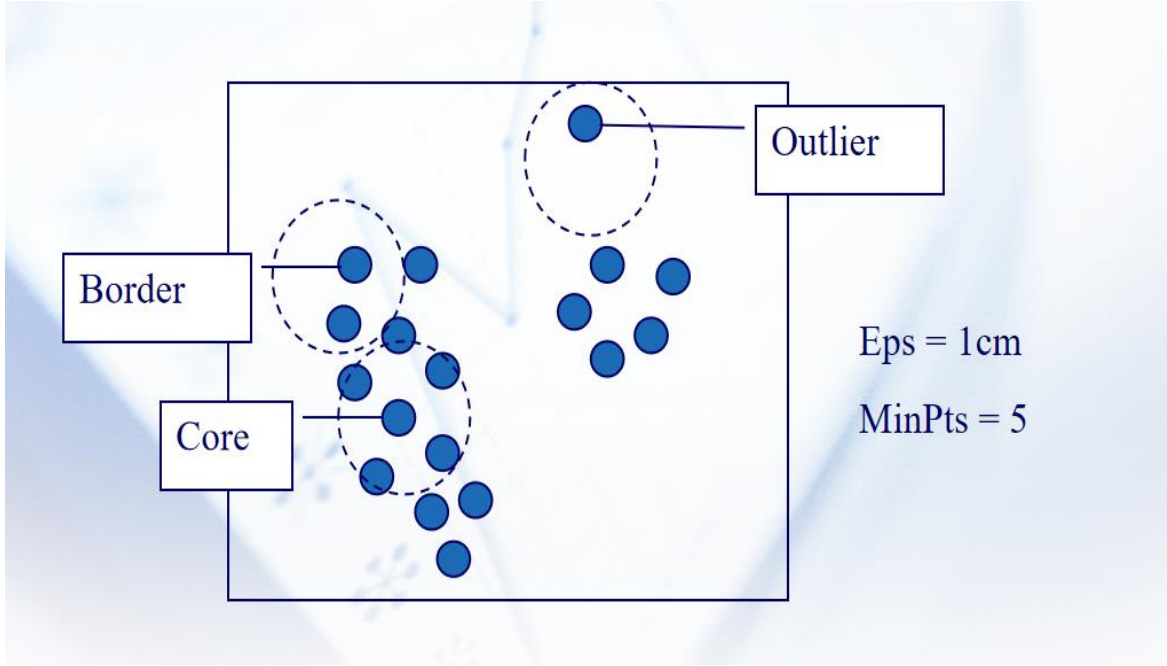
算法性能：
　　时间复杂度低，但对样本的先后顺序有一定的依赖关系，适合大数据



DIANA算法



两个重要参数

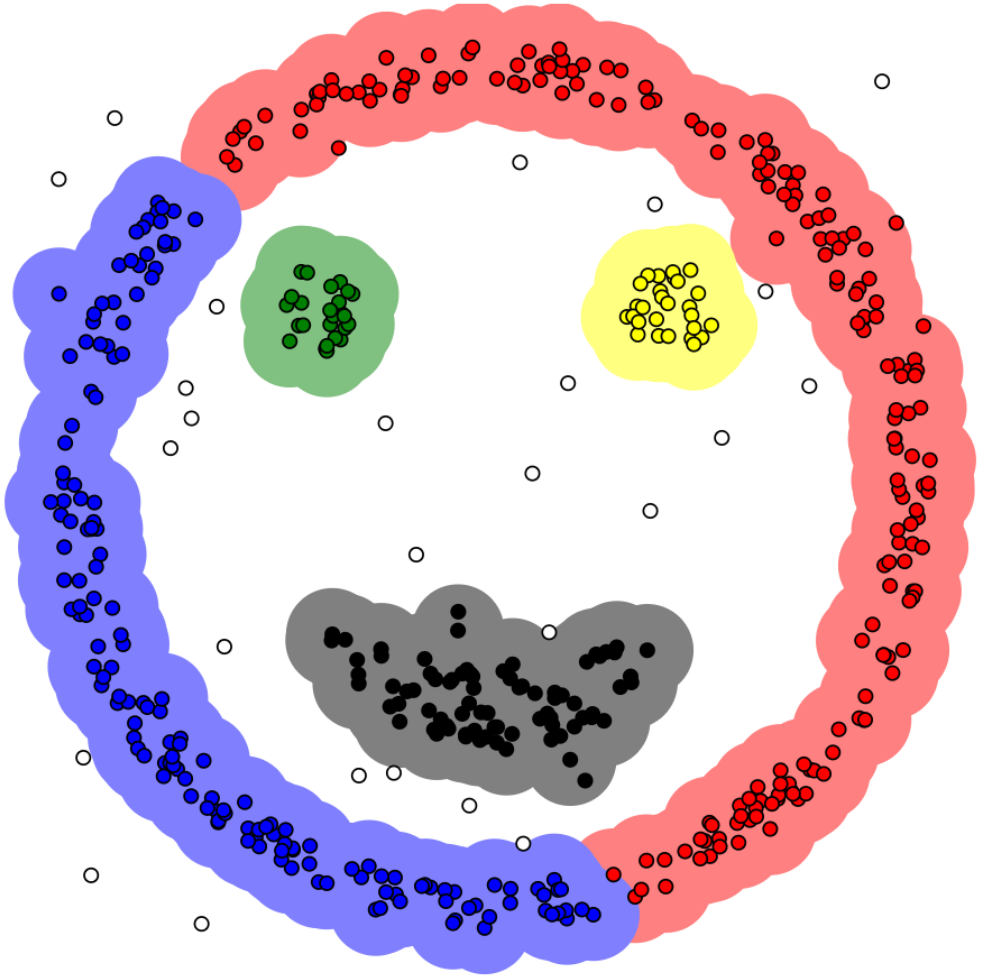


ε邻域：给定点的半径ε内的邻域

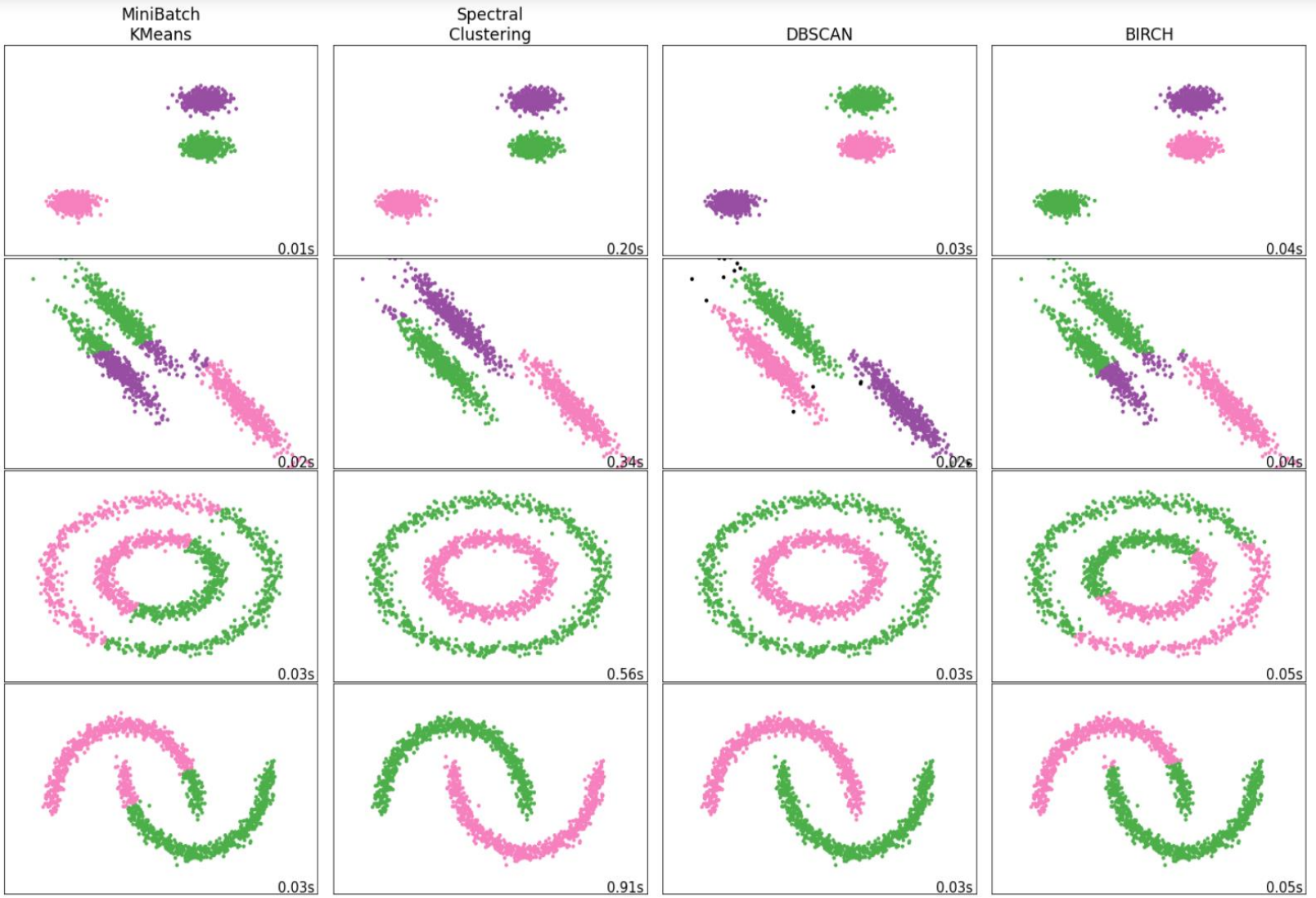
minpoints：当邻域半径ε内的点的个数大于最少点数目minpoints时，为核心点

密度可达：p1->p2->...->pn,则pn是从p1关于ε邻域和minpoints密度可达的

可以发现非凸面形状的簇，但时间复杂度为**O(n^2)**



实验对比


MiniBatch KMeans
Spectral Clustering
DBSCAN
BIRCH
0.01s
0.20s
0.03s
0.04s
0.02s
0.34s
0.02s
0.04s
0.03s
0.56s
0.03s
0.05s
0.03s
0.91s
0.03s
0.05s

性能对比

聚类算法	处理大数据的能力	处理高维数据的能力	发现任意形状簇的能力	处理噪声的能力
基于划分	较弱	强	较强	弱
基于层次	弱	较强	强	较弱
基于密度	较强	弱	强	强

Calinsk-i Harabasz(CH)指数

$$s(k) = \frac{tr(B_k)m - k}{tr(W_k)k - 1}$$

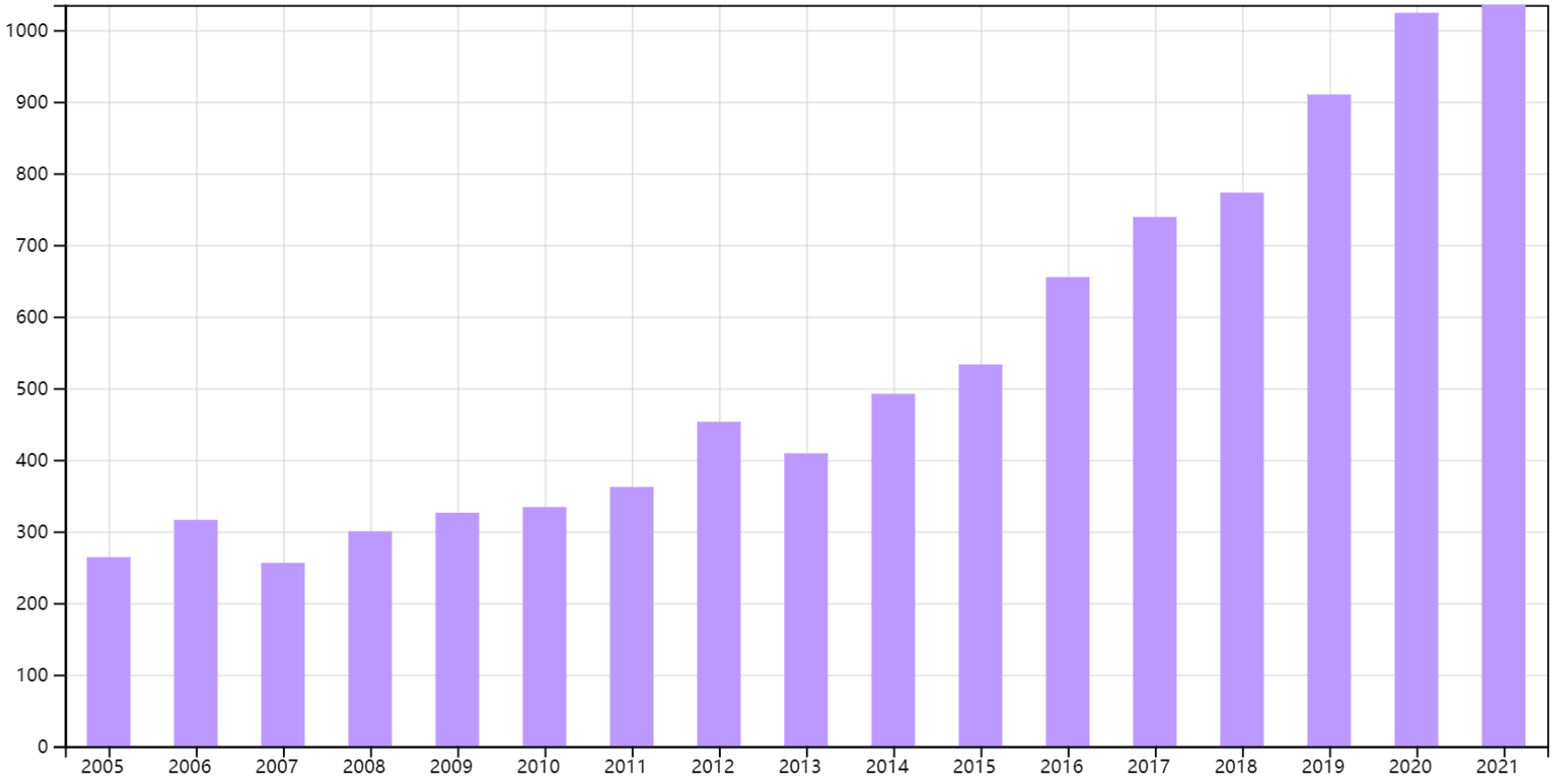
CH越大代表着类自身越紧密，类与类之间越分散，即更优的聚类结果。

Davies-Bouldin(DB)指数

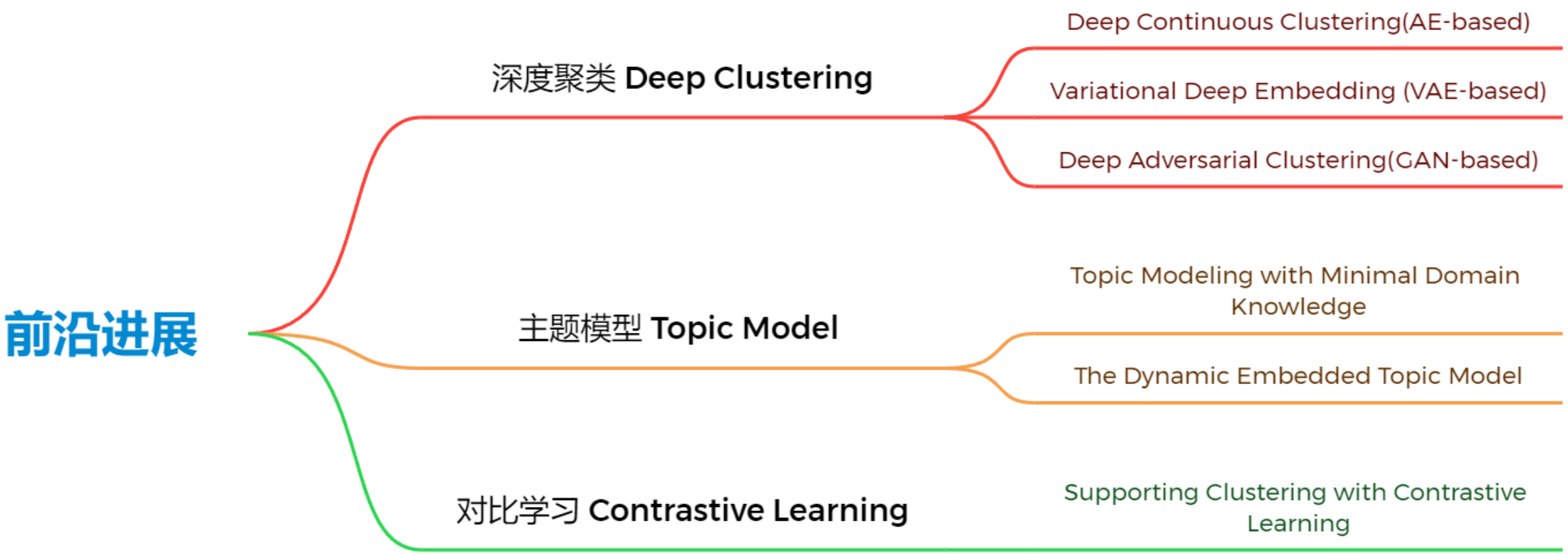
$$DB = \frac{1}{n}\sum_{i=1}^{n} max(j \neq i)(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)})$$

DB越小意味着类内距离越小同时类间距离越大

4 前沿进展



以text clustering 为主题的论文发表数量 ——来源 web of science


前沿进展
深度聚类 Deep Clustering
Deep Continuous Clustering(AE-based)
Variational Deep Embedding (VAE-based)
Deep Adversarial Clustering(GAN-based)
主题模型 Topic Model
Topic Modeling with Minimal Domain Knowledge
The Dynamic Embedded Topic Model
对比学习 Contrastive Learning
Supporting Clustering with Contrastive Learning

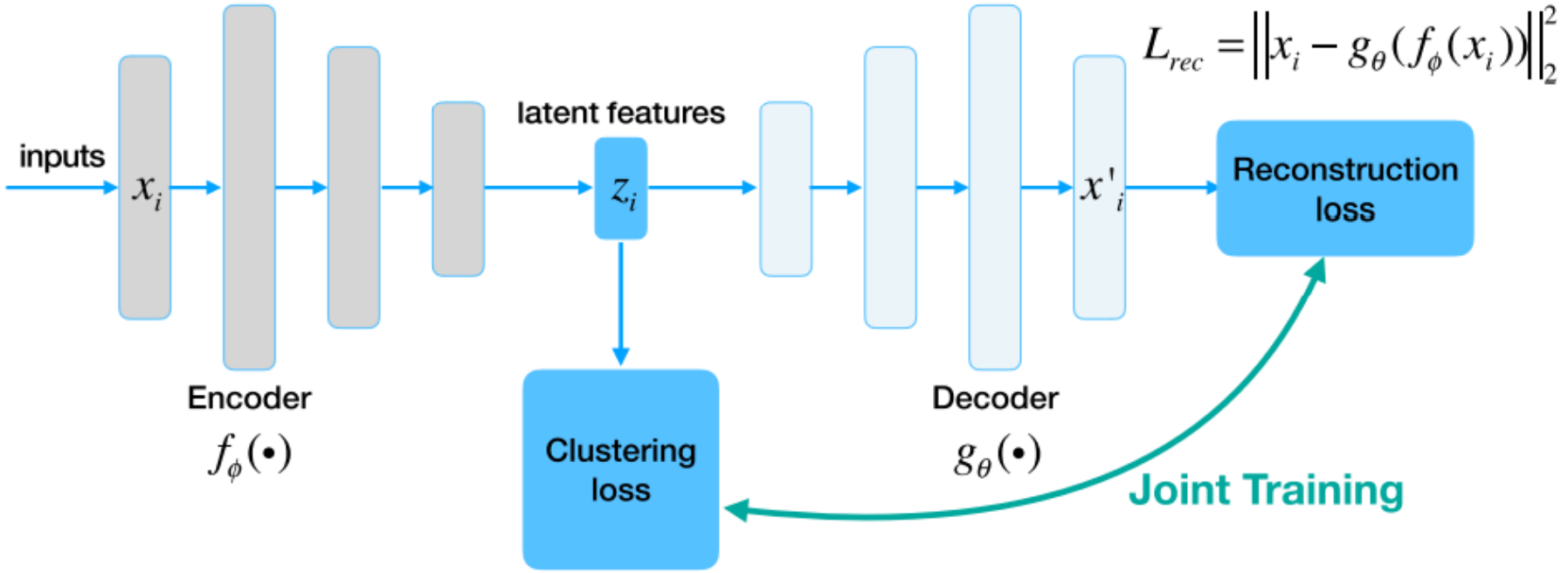


FIGURE 1. Architecture of clustering based on autoencoder. The network is trained by both clustering loss and reconstruction loss.

*资料源自A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture ERXUE MIN.ect

深度聚类 VAE-based

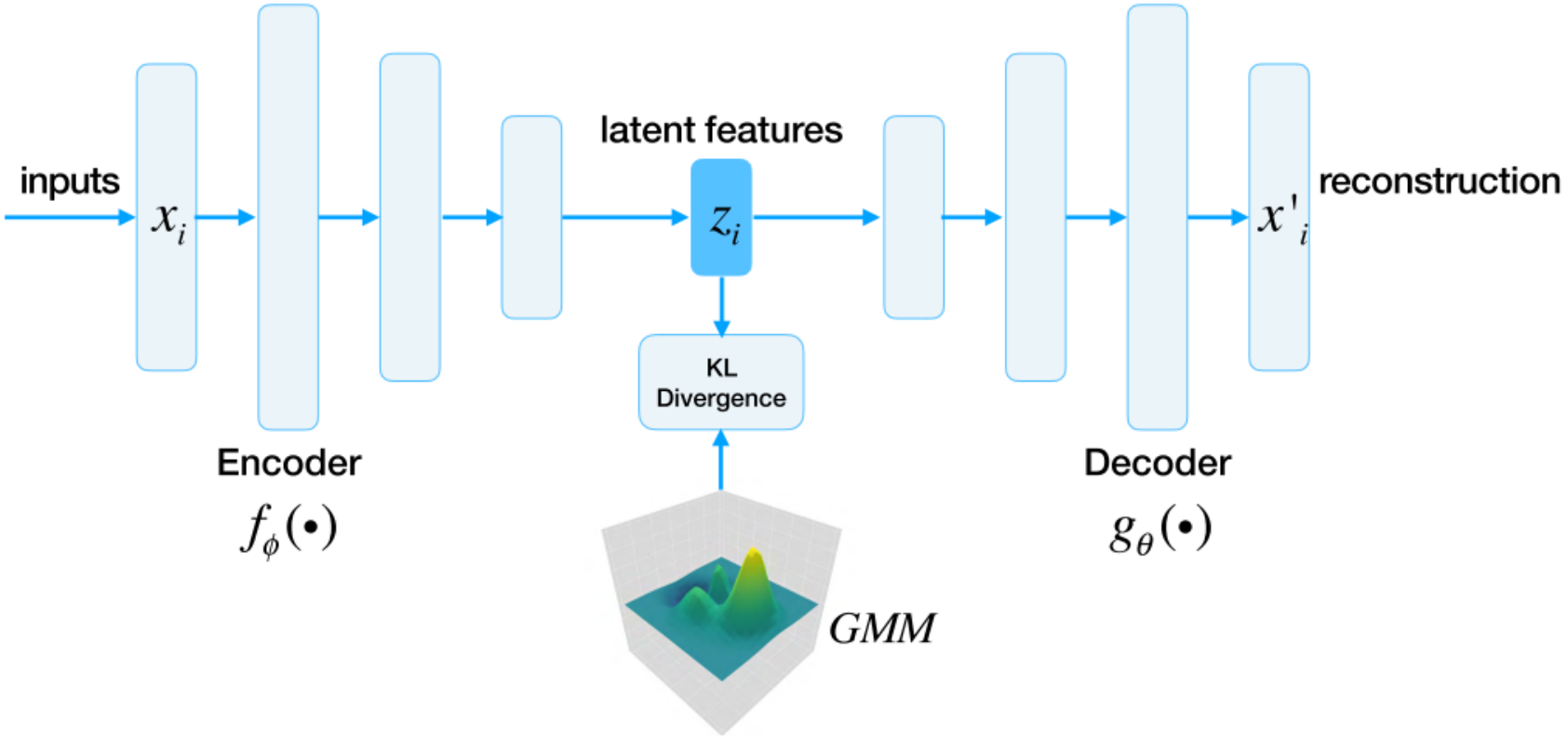


FIGURE 3. Architecture of VAE-based deep clustering algorithms. They impose a GMM priori over the latent code.

*资料源自A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture ERXUE MIN.ect

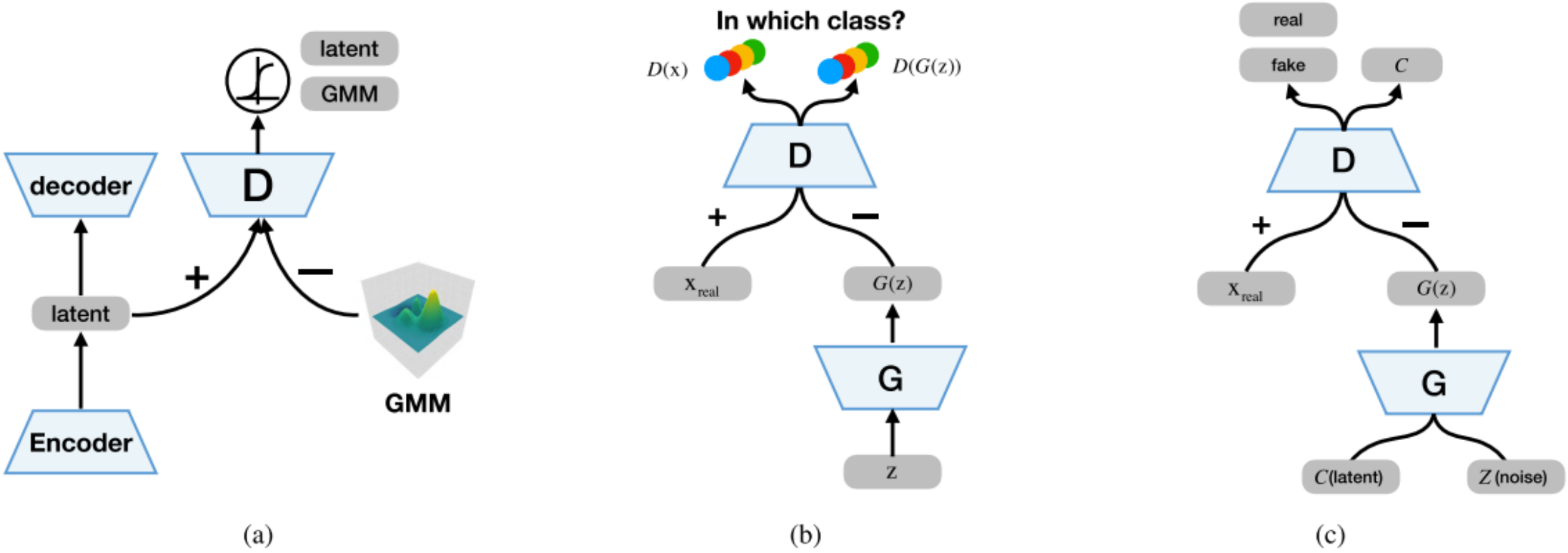


FIGURE 4. GAN-based deep clustering. (a) DAC. (b) CatGAN. (c) InfoGAN.

*资料源自A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture ERXUE MIN.ect

1. Towards K-means-friendly spaces: Simultaneous deep learning and clustering. Bo Yang et.al. (AE-based)

本文提出的DCN算法结合了自动编码器和k-means算法。DCN预先训练一个自动编码器，而后将优化重建损失和k均值损失。与其他方法相比，DCN的目标很简单，并且计算复杂度较低

2. Variational deep embedding: An unsupervised and generative approach to clustering. Z Jiang et.al. (VAE-based)

本文提出了一个基于VAE的深度无监督聚类算法VaDE，该算法结合了VAE和GMM，提出用SVGB优化方法

3. Interpretable representation learning by information maximizing generative adversarial Nets. Xi Chen et.al.(GAN-based)

本文提出的InfoGAN是一种无监督的方法，可以学习解开表示disentangled representations，也可以用于聚类。它可以解开离散和连续的潜在因素latent factors,，扩展到复杂的数据集

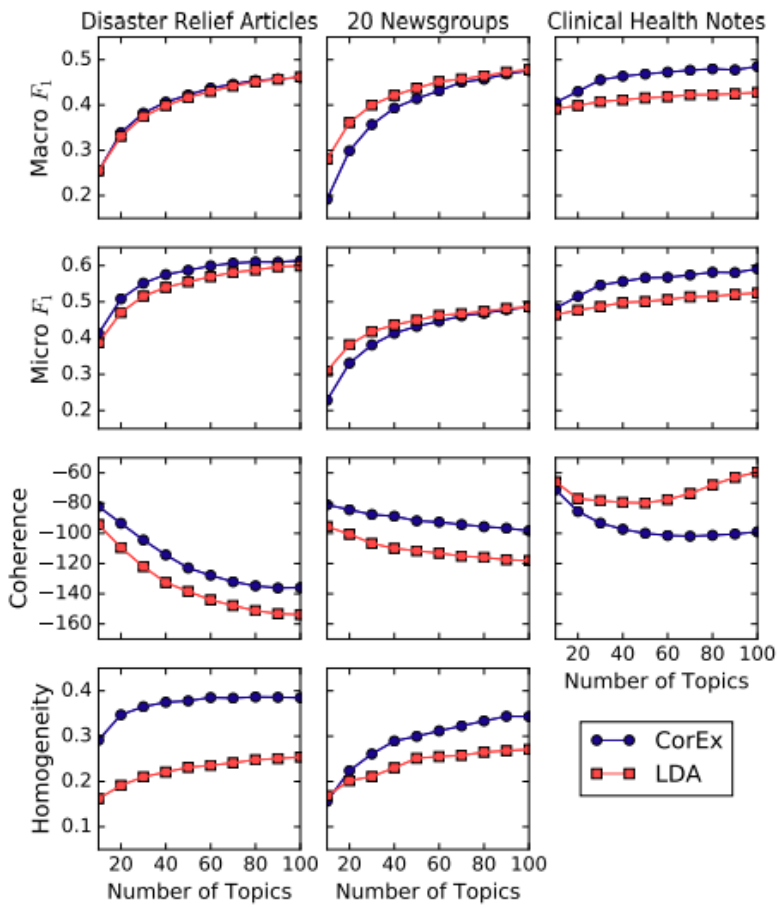
硬聚类

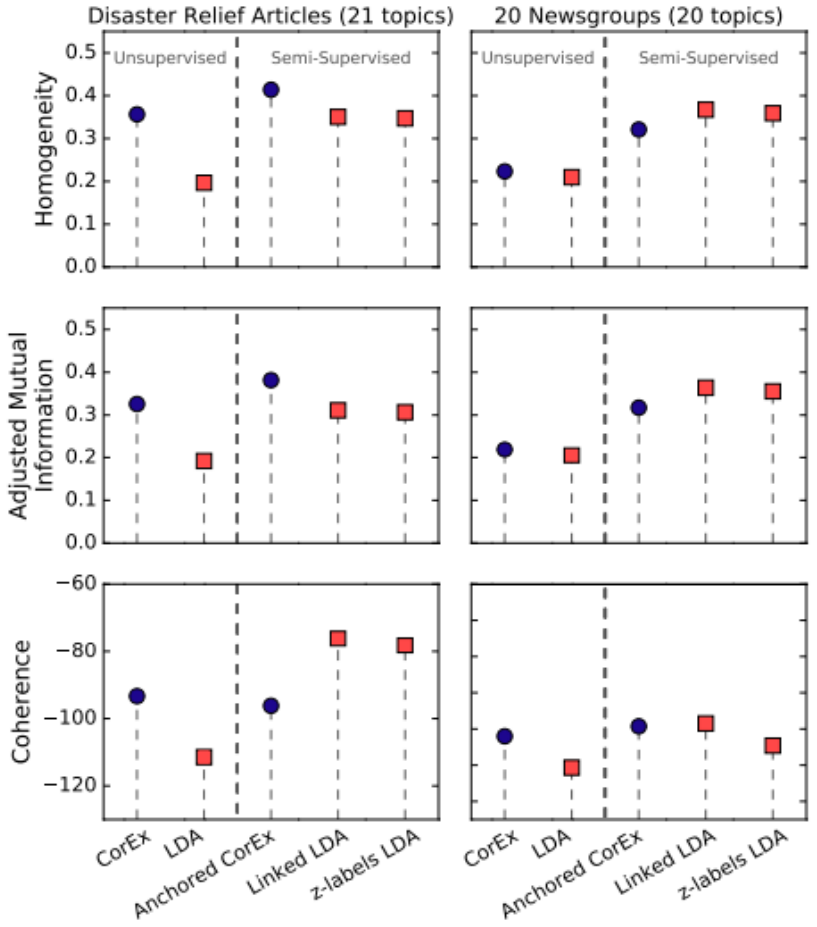
硬聚类对象存在且仅存在
唯一一个类别

软聚类

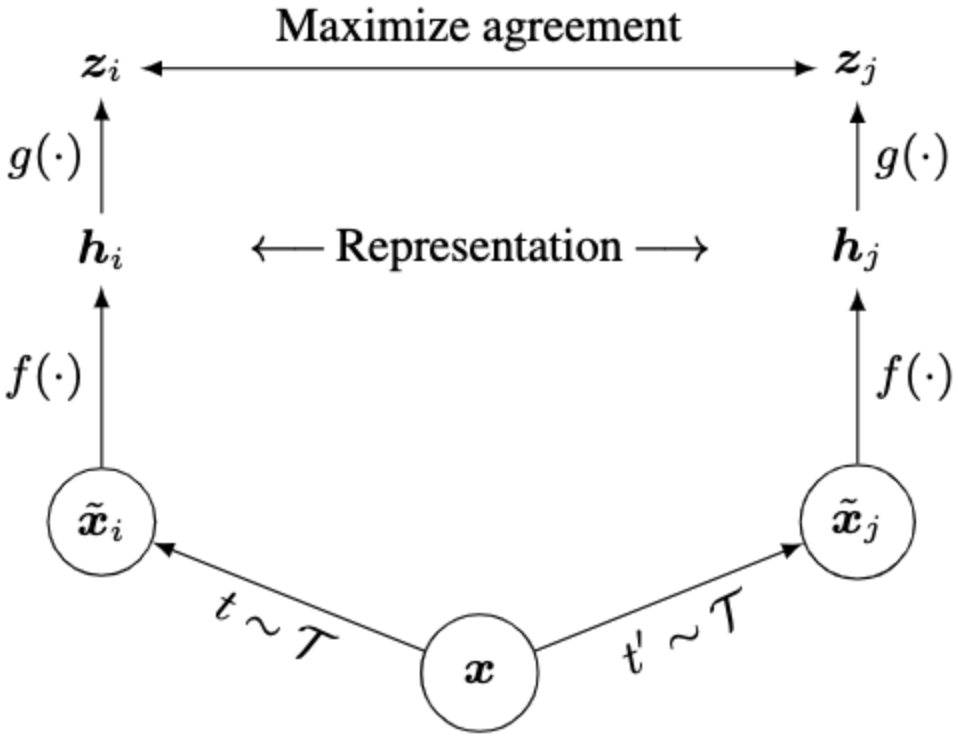
软聚类对象则可能出现在两个
或者多个类别中

掺入少许先验知识的主题模型

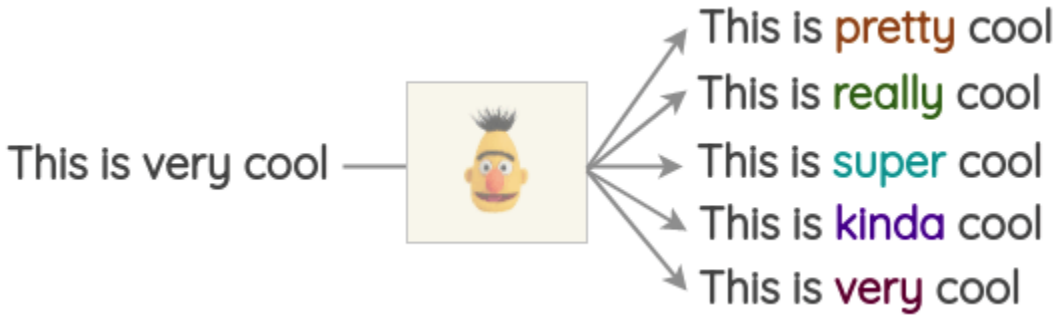




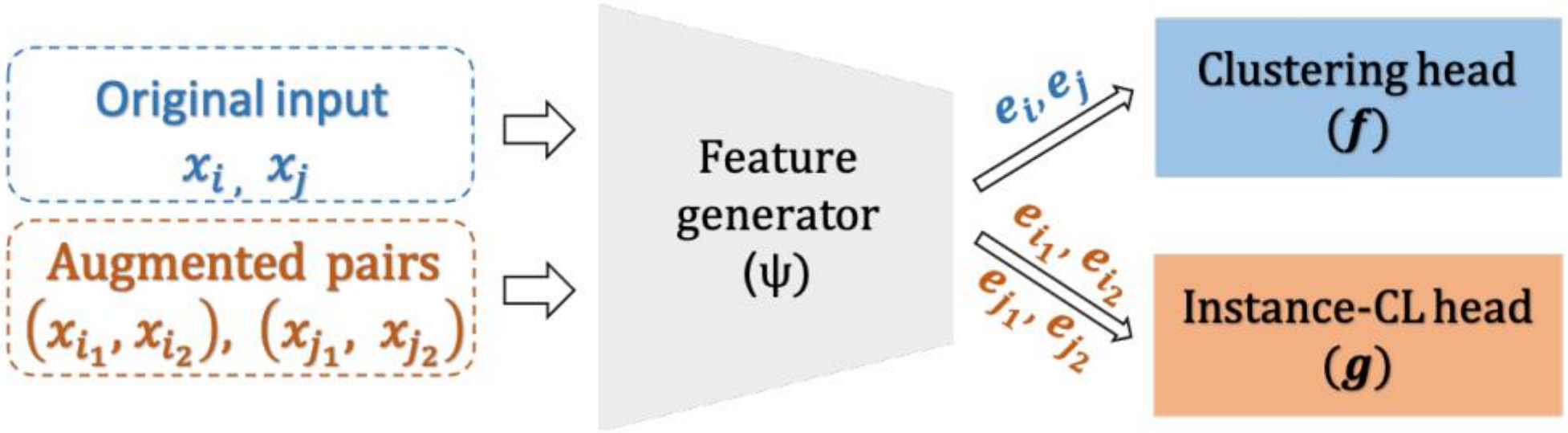
*资料源自Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge



SimCLR框架中对比学习思想的示例图



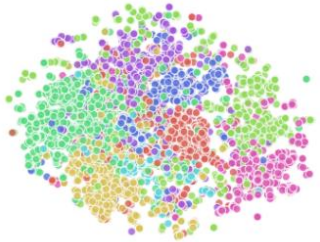
对比学习的目标是区分两个实例是否是由同一个源数据采样/增强得来，如果是，让它们在表示空间中越接近；如果不是，让它们在表示空间中远离。



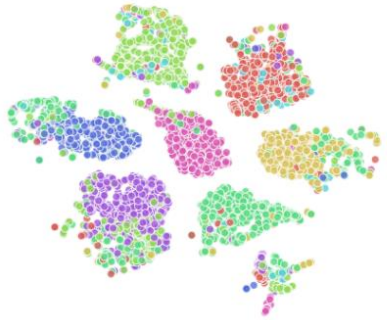
SCCL训练框架

*资料源自Supporting Clustering with Contrastive Learning

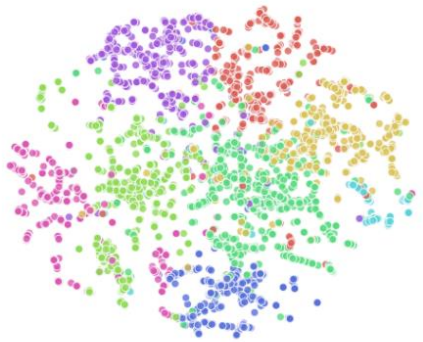
Original



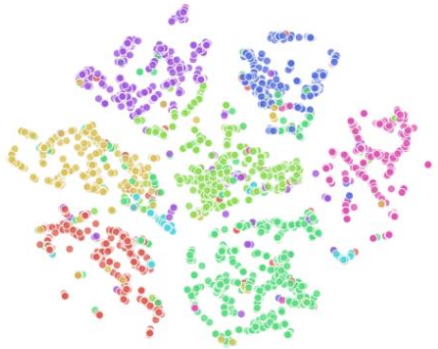
Clustering



Instance-CL



SCCL

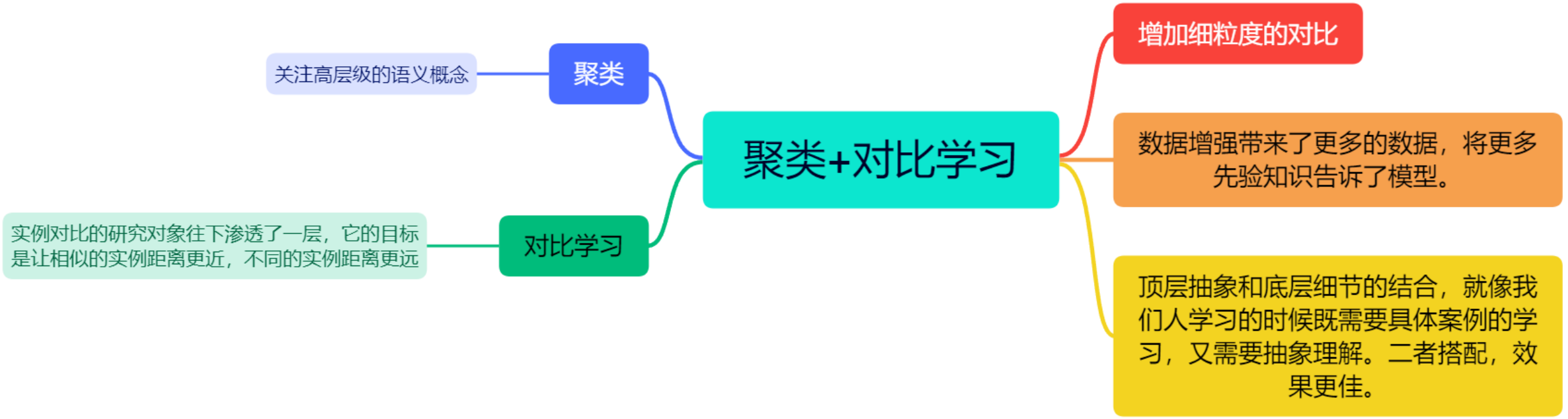


	AgNews		SearchSnippets		StackOverflow		Biomedical	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
BoW	27.6	2.6	24.3	9.3	18.5	14.0	14.3	9.2
TF-IDF	34.5	11.9	31.5	19.2	58.4	58.7	28.3	23.2
STCC	-	-	77.0	63.2	51.1	49.0	43.6	38.1
Self-Train	-	-	77.1	56.7	59.8	54.8	**54.8**	**47.1**
HAC-SD	81.8	54.6	82.7	63.8	64.8	59.5	40.1	33.5
SCCL	**88.2**	**68.2**	**85.2**	**71.1**	**75.5**	**74.5**	46.2	41.5

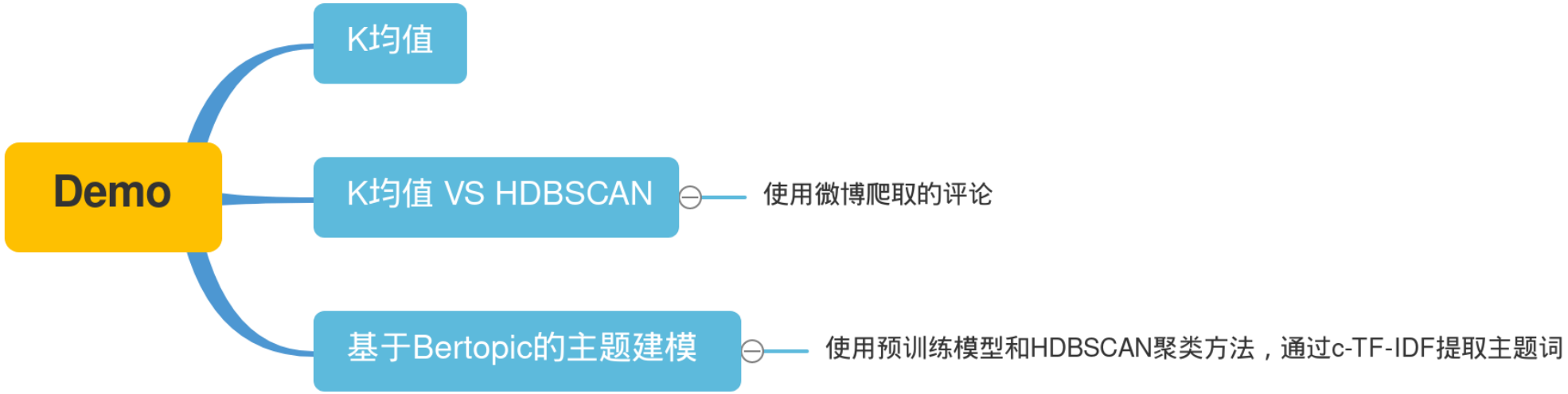
	GoogleNews-TS		GoogleNews-T		GoogleNews-S		Tweet	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
BoW	57.5	81.9	49.8	73.2	49.0	73.5	49.7	73.6
TF-IDF	68.0	88.9	58.9	79.3	61.9	83.0	57.0	80.7
STCC	-	-	-	-	-	-	-	-
Self-Train	-	-	-	-	-	-	-	-
HAC-SD	85.8	88.0	**81.8**	84.2	80.6	83.5	**89.6**	85.2
SCCL	**89.8**	**94.9**	75.8	**88.3**	**83.1**	**90.4**	78.2	**89.2**

*资料源自Supporting Clustering with Contrastive Learning

为什么聚类+对比学习可以带来如此大的提升？



5 Demo展示


Demo
K均值
K均值 VS HDBSCAN
使用微博爬取的评论
基于Bertopic的主题建模
使用预训练模型和HDBSCAN聚类方法，通过c-TF-IDF提取主题词

```python
from sentence_transformers import SentenceTransformer
from sklearn.cluster import KMeans

embedder = SentenceTransformer('all-MiniLM-L6-v2')

corpus = ['A man is eating food.',
          'A man is eating a piece of bread.',
          'A man is eating pasta.',
          'The girl is carrying a baby.',
          'The baby is carried by the woman',
          'A man is riding a horse.',
          'A man is riding a white horse on an enclosed ground.',
          'A monkey is playing drums.',
          'Someone in a gorilla costume is playing a set of drums.',
          'A cheetah is running behind its prey.',
          'A cheetah chases prey on across a field.'
          ]
corpus_embeddings = embedder.encode(corpus)

num_clusters = 5
clustering_model = KMeans(n_clusters=num_clusters)
clustering_model.fit(corpus_embeddings)
cluster_assignment = clustering_model.labels_

clustered_sentences = [[] for i in range(num_clusters)]
for sentence_id, cluster_id in enumerate(cluster_assignment):
    clustered_sentences[cluster_id].append(corpus[sentence_id])

for i, cluster in enumerate(clustered_sentences):
    print("Cluster ", i+1)
    print(cluster)
    print("")
```

```
Cluster  1
['A man is riding a horse.', 'A man is riding a white horse on an enclosed ground.']

Cluster  2
['A man is eating food.', 'A man is eating a piece of bread.', 'A man is eating pasta.']

Cluster  3
['A cheetah is running behind its prey.', 'A cheetah chases prey on across a field.']

Cluster  4
['The girl is carrying a baby.', 'The baby is carried by the woman']

Cluster  5
['A monkey is playing drums.', 'Someone in a gorilla costume is playing a set of drums.']
```

```
Cluster  1
['早睡早起痛快玩']

Cluster  2
['我再多打点音游是不是就能变成音柱了.', '这鬼灭十分的珍贵，应该让我先看', '这个酸菜十份的珍贵，应该让大家先吃，我不打扰了，我走了哈！',
'我觉得菈妮不会对9智力的笨比褪色者感兴趣的', '又想骗我玩，看看就好了，这辈子不会碰魂系列的', '所以说你们还在等什么呢，今晚8点我在沙城等
你']

Cluster  3
['不能晒太阳的你，再强大也是假的', '为什么鬼灭晚上播，因为白天有太阳', '刘华强是专门打击不良商贩的质量侠', '总有一天，全城的炊事员都要
高看我']

Cluster  4
['本作战斗确实适合体验党，但是对于喜欢深挖招式处理的玩家十分折磨，我的建议是快乐游戏别深玩，就玩帅的，怎么帅怎么来']

Cluster  5
['游戏里的你，再强大也是真的，不是假的', '充了钱，游戏里的你，再假，也是强大的']
```

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

```
<bound method DataFrame.info of                                               评论内容
0    王冰冰这个名字可真难写，倒不是笔画繁琐，只是写的时候要蘸上四分黄昏，三分月色，两分微醺，还有...
1                                             老婆老婆老婆
2                                  看来我们要众筹给调音师了，你说是吧冰酱
3                                          已经单曲循环一天了
4                                         听冰一首歌，如听一首歌
..                                                 ...
309                                             宝贝 放过我
310                                                NaN
311                                         冰冰，可分享穿搭吗？
312                                           此时无声胜有声
313

[314 rows x 1 columns]>
```

爬取了314条微博评论，只选取了评论内容部分

文本嵌入模型选择预训练的Bert模型，
由sentence_transformers库提供：

```
embedder = SentenceTransformer('all-MiniLM-L6-v2')
corpus_embeddings = embedder.encode(corpus)
db = DBSCAN(eps=0.2, min_samples=4).fit(corpus_embeddings)
```

```
embedder = SentenceTransformer('all-MiniLM-L6-v2')
num_clusters = 8
corpus_embeddings = embedder.encode(corpus)
db = KMeans(n_clusters=num_clusters).fit(corpus_embeddings)
```

HDBSCAN结果

	评论内容	cluster
1	老婆老婆老婆	0
10	图片评论	0
17	老婆你说啥就是啥	0
25	好听好听	0
26	送你🍎啊冰冰	0
36	冰冰	0
42	老婆老婆老婆老婆老婆	0
50	老婆	0
54	啊对对对对对	0
61	老婆老婆老婆	0
70	老婆	0
78	你说是就是吧	0
79	冰冰老婆	0
81	啊对对对	0
82	听到了	0
83	网页链接	0
84	又要辛苦修音师傅了	0
93	图片评论	0
97	喵咪咪喵咪咪喵咪咪	0
98	那当然了	0

	评论内容	cluster
96	是	3
121	🙈🙈🙈	3
196	哇	3
261	宝	3
281	是	3

	评论内容	cluster
71	好美哦	2
143	好美啊	2
178	头饰美	2
303	老婆好美	2

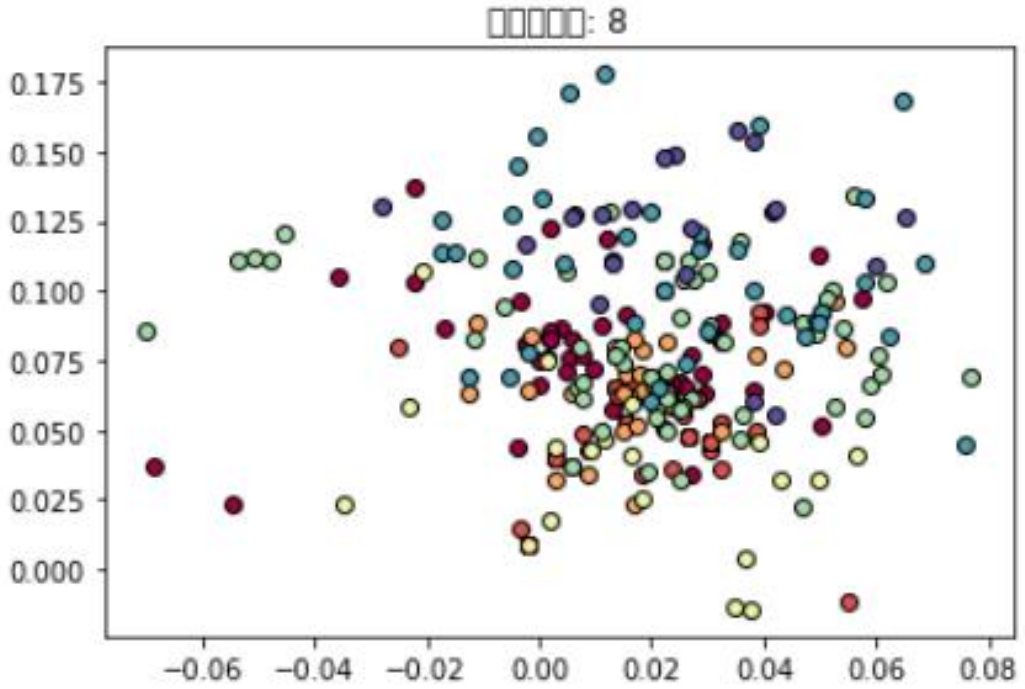
	评论内容	cluster
123	老婆给我亲亲	4
130	我听到了	4
135	老婆我爱你	4
151	我来迟了	4
247	我感受到了爱情	4
272	冰冰我爱你	4
309	宝贝 放过我	4

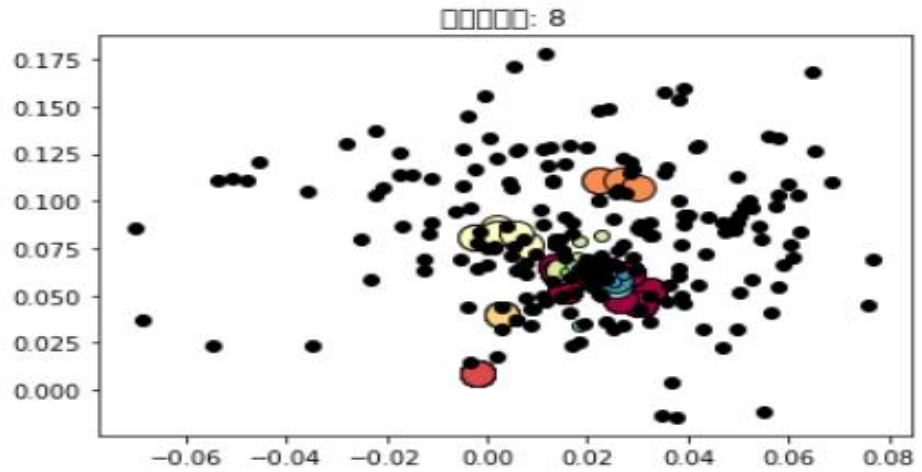
K均值结果

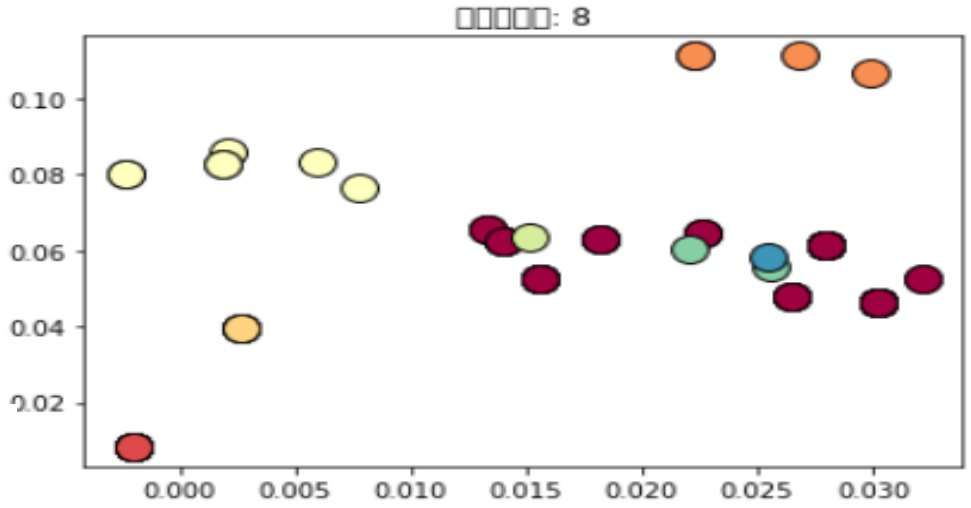
	评论内容	cluster
1	老婆老婆老婆	1
9	谢谢，照片很动听，声音很好看	1
10	图片评论	1
14	是是是是是是是是是是是是是是是是是是是是是是是是是是是是是	1
17	老婆你说啥就是啥	1
25	好听好听	1
26	送你🍎啊冰冰	1
31	姐姐，爱你	1
36	冰冰	1
42	老婆老婆老婆老婆老婆	1
43	哇，好好听哦	1
50	老婆	1
54	啊对对对对对	1
55	辛苦了，调音师。	1
61	老婆老婆老婆	1
63	哦？是嘛	1
70	老婆	1
78	你说是就是吧	1
79	冰冰老婆	1
81	啊对对对	1

	评论内容	cluster
2	看来我们要众筹给调音师了，你说是吧冰酱	0
7	冰冰，明天要考研了，我直接裸考进考场，你能祝福我吗	0
20	我给我耳朵买保险了	0
22	你。。你要这么说的话。。那我也可以勉强摘下耳机	0
28	不行我要看视频，不然我感觉不到	0
37	老婆老婆！！！！	0
40	中间有一句唱的不太好	0
46	是的呢（然而我只看到了蝴蝶结）	0
51	我只感觉到一只鸽子的叫声	0
56	我的脑海里还回响着"有点饿啊，没关系～"	0
59	我们是冠军！	0
60	发个视频我就说这是天籁之音不然我我就不说	0
66	老婆到我枕头边上唱	0
90	难道就我关注了大红蝴蝶结🎀吗	0
91	你们都没老婆的？非要叫我老婆老婆	0
110	不愧是我老婆	0
122	一般毕竟睡前你常唱给我听	0
123	老婆给我亲亲	0
130	我听到了	0
131	我老婆好美	0

5.3 HDBSCAN VS K均值







上为K均值，右为HDBSCAN

聚类数： 8
噪点数： 205

基于Bertopic的主题建模——基本流程

1 **数据集获取**

使用东京2020推文

2 **文本预处理**

未做详细数据清理，仅选取部分推文，然后字符化处理

3 **Bert文本嵌入**

使用Bert作为预训练嵌入模型，获得文本的向量表示

4 **聚类**

通过UMAP和HDBSCAN，将文本嵌入进行聚类，语义相近的语句将聚集成簇群

5 **提取主题词**

用c-TF-IDF提取主题词

5.4 基于Bertopic的主题建模

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

数据集使用2020东京奥运会推文（推特）：

```
['id', 'user_name', 'user_location', 'user_description', 'user_created', 'user_followers', 'user_friends', 'user_favourites', 'user_verified', 'date', 'text', 'hashtags', 'source', 'retweets', 'favorites', 'is_retweet']
```

	source	retweets	favorites	is_retweet
0	Twitter for Android	0.0	0.0	False
1	Twitter for Android	0.0	0.0	False
2	Twitter for Android	0.0	1.0	False
3	Twitter Web App	1.0	0.0	False
4	Twitter for iPhone	0.0	0.0	False
...	...	...	...	...
160543	Twitter Web App	7.0	38.0	False
160544	Twitter Web App	7.0	11.0	False
160545	Twitter for iPhone	78.0	355.0	False
160546	Twitter for iPhone	0.0	0.0	False
160547	Twitter Web App	0.0	0.0	False

[160548 rows x 16 columns]>

	user_description
0	Trying to be mediocre in many things
1	Indian weightlifter 48 kg category. Champion🏆
2	All breaking news related to Financial Market....
3	Official International Hockey Federation Twitt...
4	Football & Tennis Coach
...	...
160543	The Belgian National Team - Women's basketball
160544	Keep updated with all the latest news from the...
160545	CBC News/Olympics Reporter. Based in Toronto. ...
160546	🗼*Tokyo JAPAN* JP ✂ 表参道 ⋆ 原宿 ✂ *Hair stylist* ✂...
160547	Powered by AI fake news filtering technology, ...

总共16万条推文，16个属性，数据内容基本是英文

5.4 流程

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

为了便于调试，选取前六万条推文作为数据，模型使用Bertopic中封装好的bert预训练模型

In [6]:
```
# create model
model = BERTopic(verbose=True,nr_topics="auto",language="multilingual")
#convert to list
docs = doc.text.to_list()
data=[str(i) for i in docs]
```

Out[6]: str

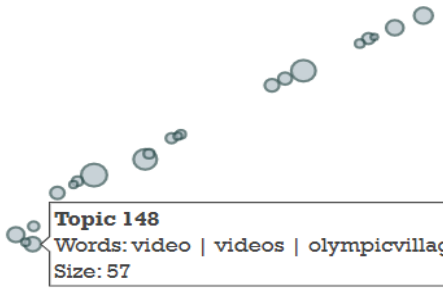
In [7]:
```
topics, probabilities = model.fit_transform(data)
```

```
Batches: 100%|██████████████████████████████████████████████████████████████████████████████| 1875/1875 [34:25<00:00,  1.10s/it]
2022-03-22 17:19:51,827 - BERTopic - Transformed documents to Embeddings
2022-03-22 17:20:30,764 - BERTopic - Reduced dimensionality with UMAP
2022-03-22 17:20:35,023 - BERTopic - Clustered UMAP embeddings with HDBSCAN
2022-03-22 17:21:06,055 - BERTopic - Reduced number of topics from 707 to 568
```

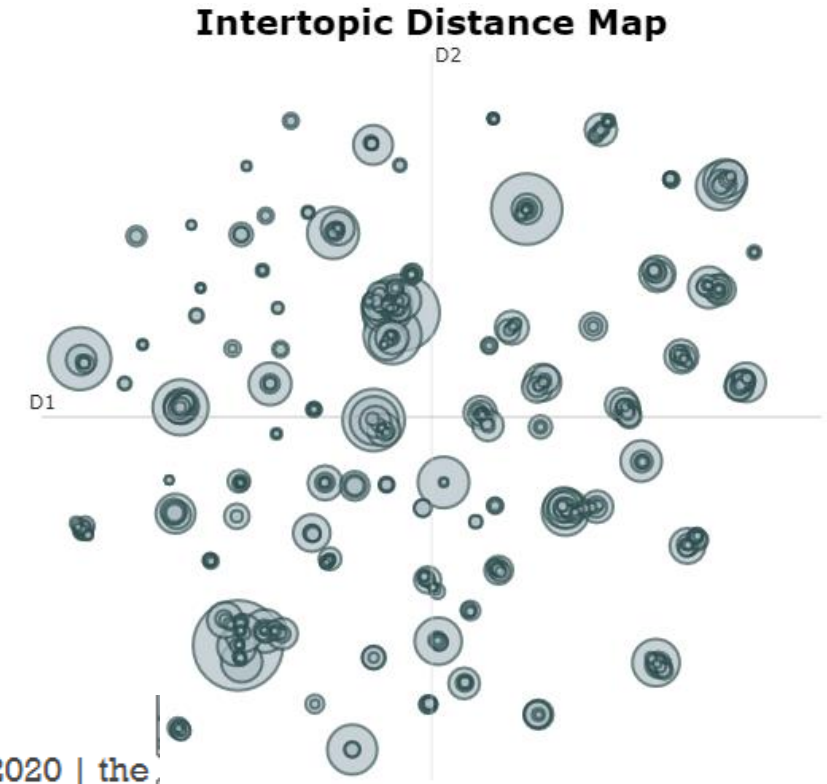
模型使用Bert进行特征提取，使用UMAP方法降维，然后使用HDBSCAN进行聚类，主题参数选择auto，模型根据相似程度把原来的707个主题减少到了568个

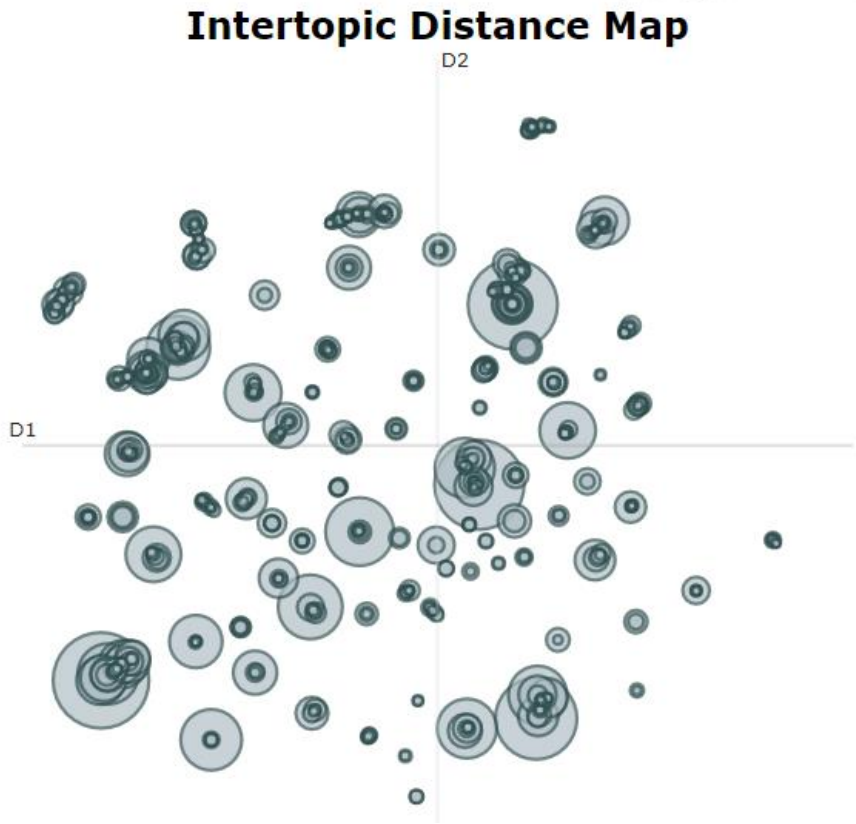
5.4 实验结果分析

在语言的参数选择上，使用了英语和多语言（支持50多种语言），左图表示英语的聚类结果，右图为多语言的聚类结果



Topic 148
Words: video | videos | olympicvillage | youtube | music
Size: 57





Topic 26
Words: swimming | medley | 400m | tokyo2020 | the
Size: 473

Topic 0
Words: mirabaichanu | weightlifting | medal | 49kg | silver
Size: 1973

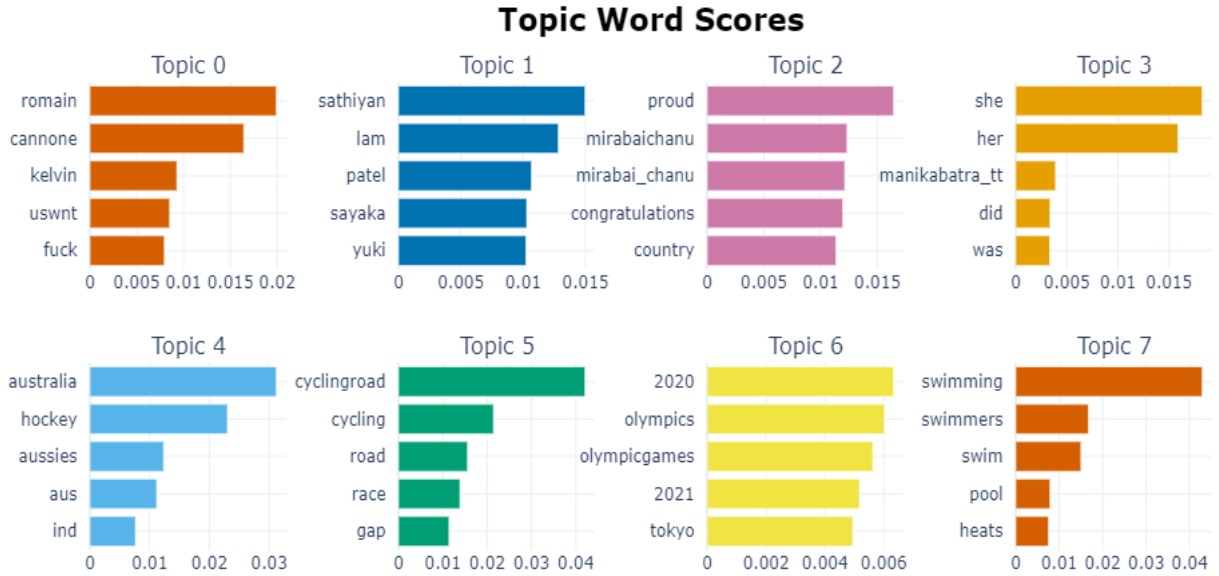
Topic 1
Words: olympicgames | olympics | olympics2021 | watching | sports
Size: 1045

5.4 实验结果分析

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

每个簇的主题词可视化

英语



多语种