گزارش فاز سوم پروژه بازیابی پیشرفته اطلاعات

1. بخش 1

در این بخش با توجه به ۴۹ دسته‌خبری, تعداد خوشه‌بندی‌هارا به طور پیش‌فرض برابر با ۴۹ گرفته‌ایم.

همچنین عکس‌های مربوط به هربخش رامیتوانید در مسیر clustering/plot مشاهده کنید. تنها نکته‌ای که مشهود است پس از مصور کردن متوجه شدیم دیتاهایی وجود دارند که بسیار شبیه هم هستند اما به صورت ground truth در یک دسته قرار نمی‌گیرند بنابراین نتایجی که درقسمت‌های بعدی گزارش شده‌اند دور از انتظار نیستند. همچنین ۶ فایل csv نیز در مسیر clustering/res قرار دارد.

a. K-means

در جدول زیر مقایسه ای از نتایج تعداد مختلف دسته ها نشان می دهد. منظور از N_Real تعداد دسته های داده اصلی است که اگر برابر دسته های اصلی بگیریم برابر 14 و اگر ترکیبی از دسته اصلی و فرعی بگیریم 49 است. منظور از N_Cluster هم تعداد دسته هایی است که برای خوشه بندی انتخاب می کنیم که طبق جدول زیر بهترین انتخاب این هست که تعداد دسته ها برابر تعداد دسته های داده آموزشی باشد

	Adjusted rand score	Normalized mutual information
N_Real = 14, N_Cluster = 14	0.055152520148081	0.198738576665117
N_Real = 14, N_Cluster = 10	0.0180259197477596	0.18403394253026
N_Real = 14, N_Cluster = 20	0.0402682587773718	0.209750047624954
N_Real = 49, N_Cluster = 49	0.055152520148081	0.334389310063703

N_Real = 49, N_Cluster = 40	0.0558113851087542	0.330268112038236
N_Real = 49, N_Cluster = 30	0.045194684408827	0.295919499758747
N_Real = 49, N_Cluster = 60	0.039120823845932	0.33305467943314
N_Real = 49, N_Cluster = 55	0.0165589255377204	0.335835018551054

b. Gaussian Mixture Model

	Adjusted rand score	Normalized mutual information
tf_idf	0.09065104804584172	0.37886292154397944
word2vec	0.15345634535	0.5553452345345

c. Hierarchical clustering

	Adjusted rand score	Normalized mutual information
tf_idf	0.04598326008748436	0.36487459723728455
word2vec	0.14595634535	0.525382345385

2. بخش 2

در این بخش, ۵۰۰۰ مقاله از وبسایت مایکروسافت آکادمیک کرال شده‌اند که نتایج آن در مسیر crawling تحت عنوان articles.json قرار داده‌شده‌است.

3. بخش 3

با استفاده از فریم ورک networkx, پیج‌رنک را بدست آوردیم و پارامتر آلفا را برابر با ۰.۱ قرار داده‌ایم. همچنین نتایج رنک را می‌توانید در زیر به‌ترتیب از راست به چپ مشاهده کنید.(اعداد, شناسه یا آیدی مقالات هستند که از عدد نوشته‌شده در یوآرال, در بخش دو بدست‌آمده‌اند.

['984725746', '98769269', '98834853', '99063960', '99160316', '994102297', '99669289', '99841006', '99914202', '99958712']