گزارش فاز سوم پروژه بازیابی پیشرفته اطلاعات

1. بخش 1

در این بخش با توجه به **تست الگوریتم‌ها در بخش ۴** با پارامترهای مختلف, بهترین نتیجه در تمامی حالت‌ها به ازای ۴۹ دسته و ماکزیمم ایتریشن ۵۰ بدست‌می‌آید.

همچنین مصورسازی‌های 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.

در تمامی الگوریتم‌ها, word2vec عملکرد بهتری نسبت به tf_idf دارد.

a. K-means

برای این روش از کتابخانه sklearn.cluster.KMeans استفاده کردیم.

	Adjusted rand score	Normalized mutual information
tf_idf	0.08656623020483799	0.39023725454549585
word2vec	0.16102719953710218	0.44608855902296796

b. Gaussian Mixture Model

	Adjusted rand score	Normalized mutual information
tf_idf	0.09065104804584172	0.37886292154397944
word2vec	0.15345634535	0.42973748547563273

c. Hierarchical clustering

برای این مدل از کتابخانه‌ی sklearn.clustering استفاده کردیم.

	Adjusted rand score	Normalized mutual information
tf_idf	0.039362930804983716	0.3634558785411254
word2vec	0.16246148840790947	0.4363632777815437

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

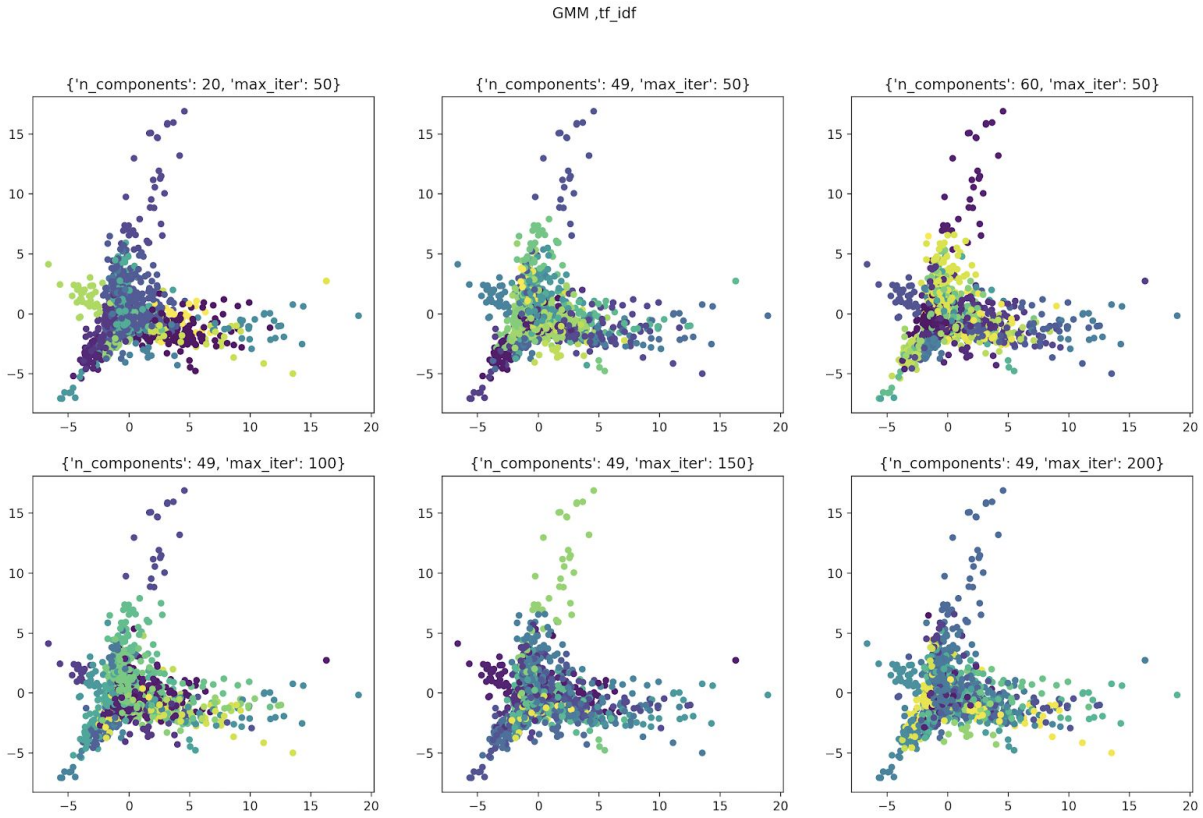
با استفاده از فریم ورک networkx, پیج‌رنک را بدست آوردیم و پارامتر آلفا را برابر با ۰.۱ قرار داده‌ایم. همچنین نتایج رنک را می‌توانید در زیر به‌ترتیب از راست به چپ مشاهده کنید.(اعداد, شناسه یا آیدی مقالات هستند که از عدد نوشته‌شده در یوآرال, در بخش دو بدست‌آمده‌اند.

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

4. تست پارامترهای مختلف

برای مشاهده‌ی نتایج مختلف دسته‌بندی‌های مختلف و تعداد iteration های متفاوت ، در فایل test_module آرایه‌ای از پارامترها برای الگوریتم های بخش اول قرار دادیم . نتایج ۶ مورد از این تست ها در تصاویر زیر مشاهده می‌شود:

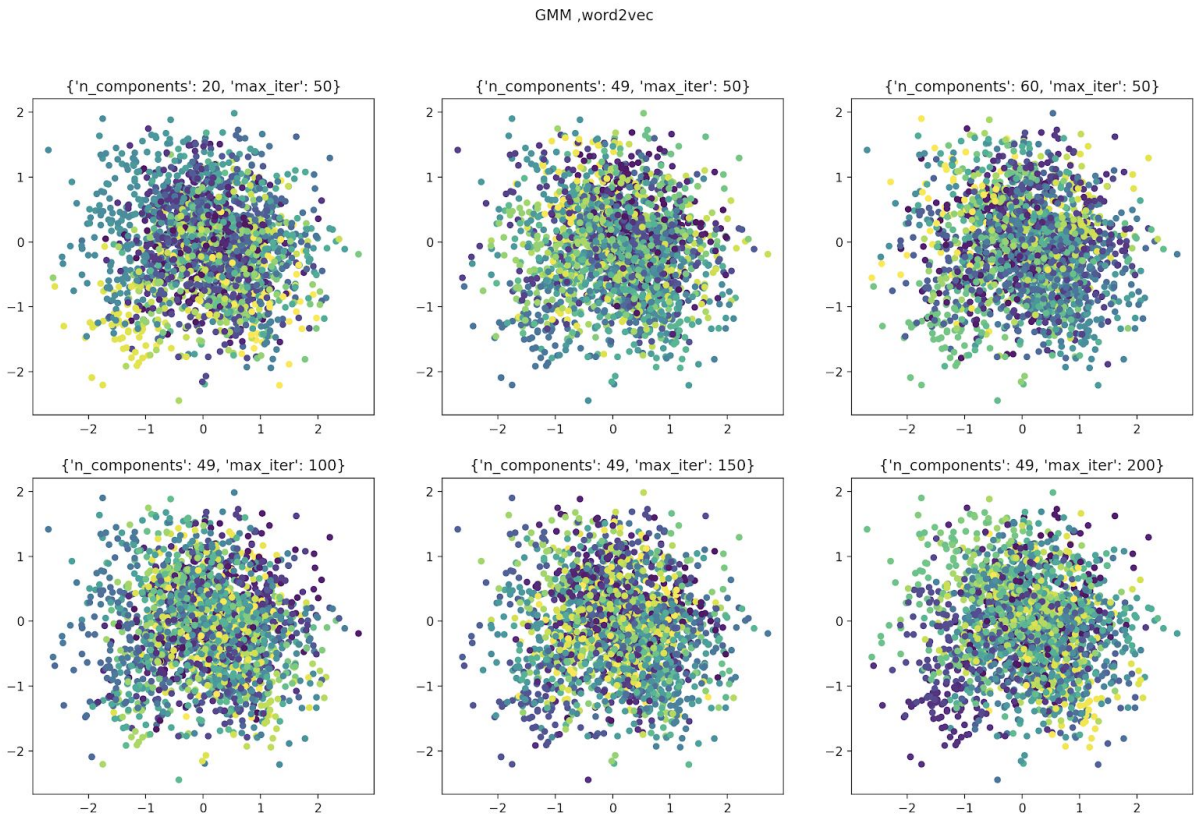
الگوریتم GMM



معیارها (از بالا سمت چپ به راست):

GMM tf_idf	Adjusted rand score	Normalized mutual information
	0.07981605125795813	0.2880824527939654
	0.12666534207667476	0.38805921806349053

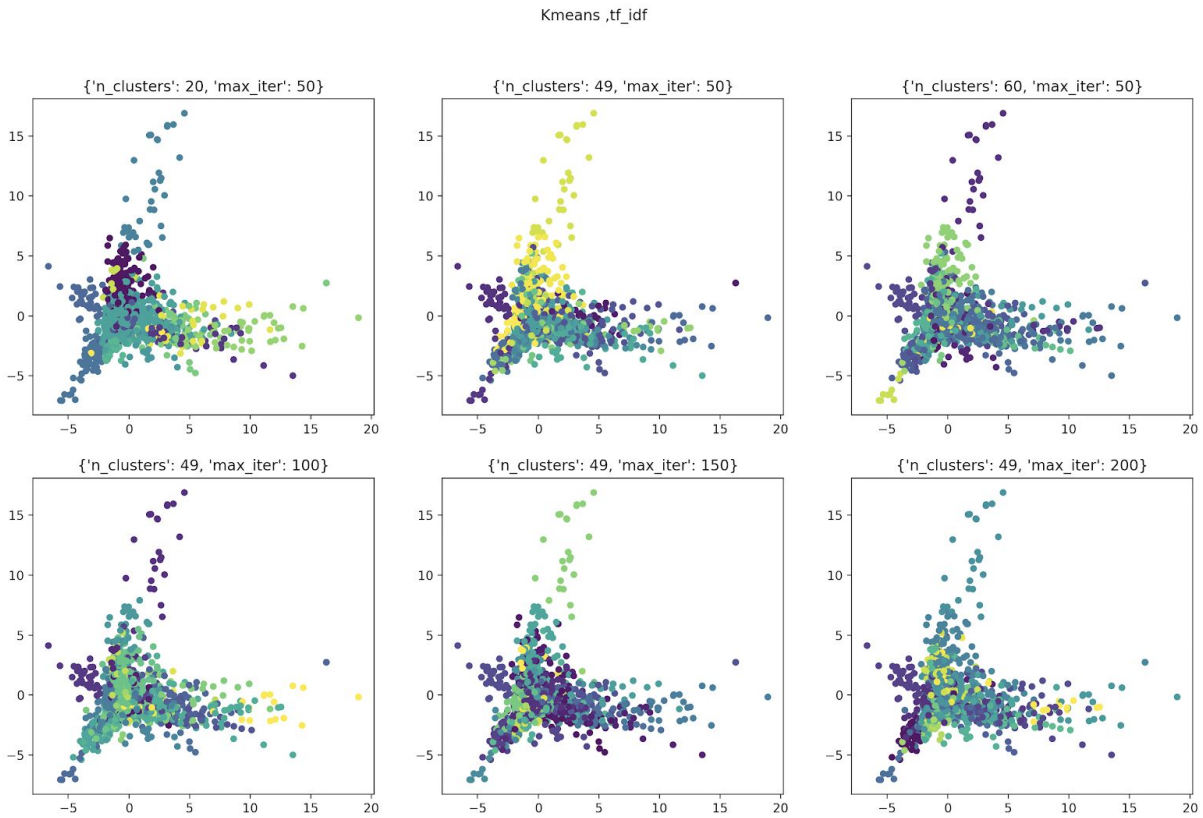
	0.11075987442777095	0.38829311563096147
	0.07611673937805404	0.3512061750361027
	0.10408444636344605	0.38718118155649034
	0.11466941953047945	0.35552647559368644



GMM word2vec	Adjusted rand score	Normalized mutual information
	0.18683789563941994	0.4097712660036656

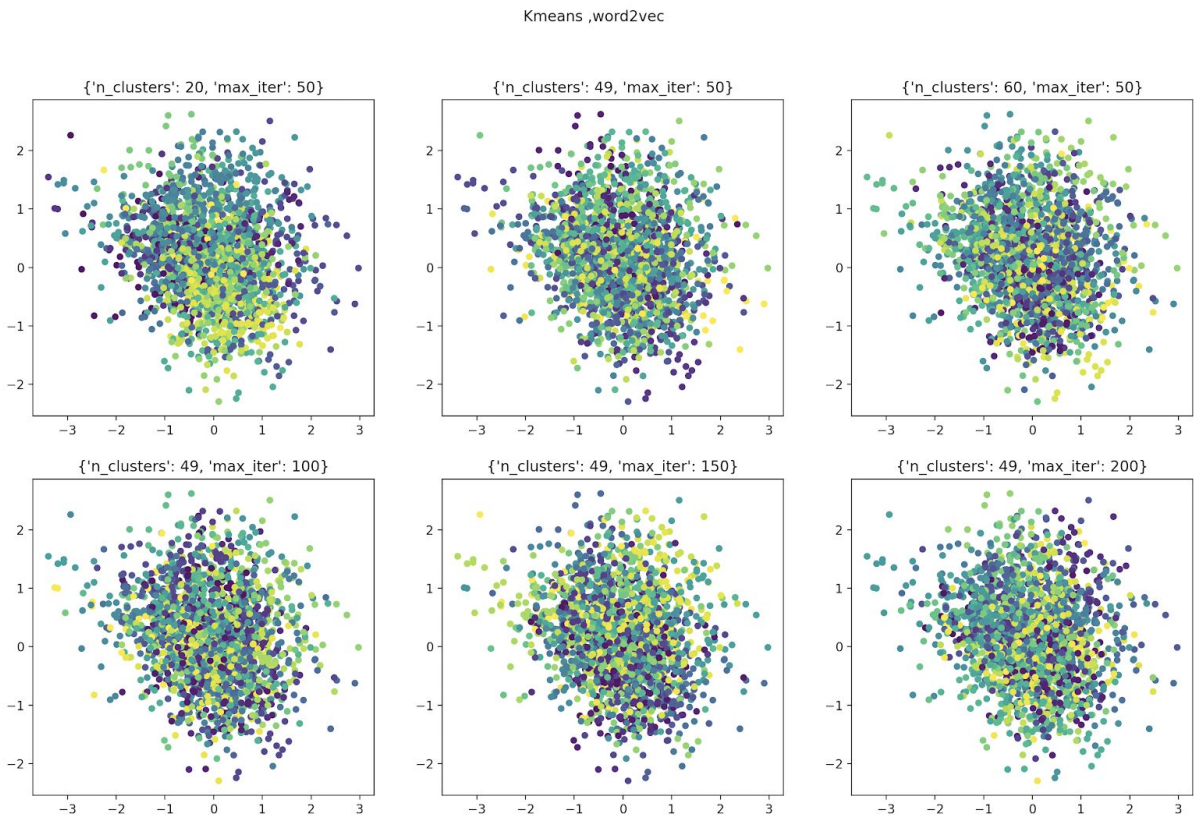
	0.16232145451504404	0.44243702491092507
	0.13414221904124238	0.4443538073285943
	0.15746854585840253	0.44352949085799526
	0.14063170207367495	0.43649105905924623
	0.1591130518763448	0.4431483462038665

الگوریتم KMeans



KMeans	Adjusted rand score	Normalized mutual information

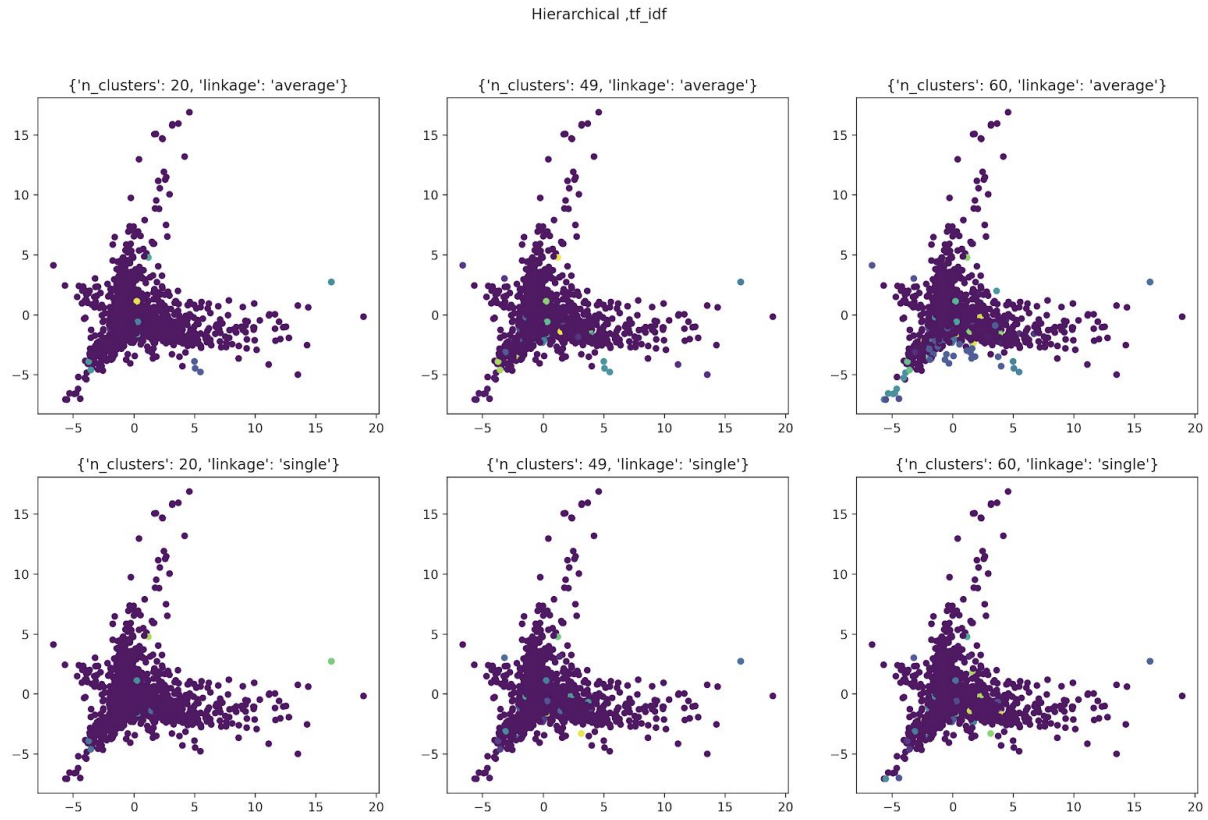
tf_idf	0.04030508743009567	0.3125454682767017
	0.08809752625220649	0.3718502494072461
	0.05465827122017849	0.3849591674674551
	0.051549604765273954	0.3738284496584375
	0.057611100055928954	0.3762466118733981
	0.0791573752262839	0.37821271302264736



KMeans	Adjusted rand score	Normalized mutual information

word2vec	0.2182387483330724	0.40699804295821546
	0.14774503575075582	0.4366328275024856
	0.12939048636364828	0.44418392384072947
	0.13718321824677962	0.4347416438274418
	0.15259585765495542	0.43539011870677075
	0.15909946846509856	0.44124693533326065

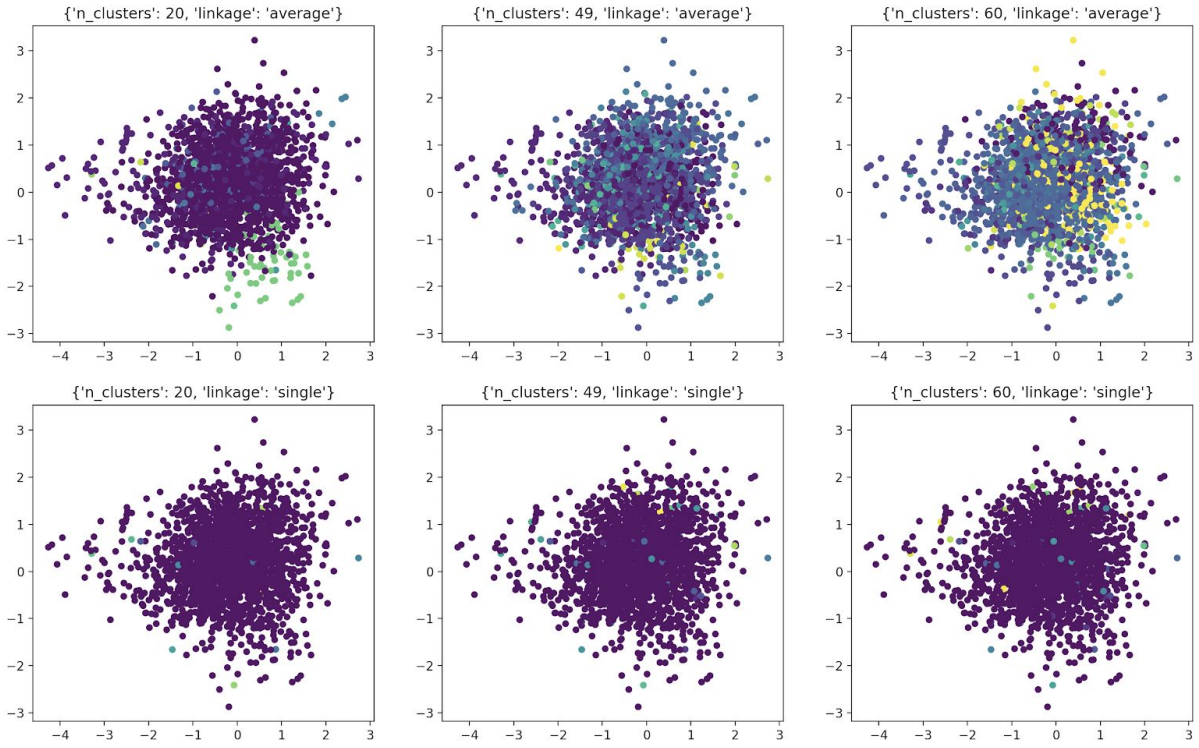
الگوریتم Hierarchical



Hierarchical	Adjusted rand score	Normalized mutual information

tf_idf	0.0002679730617417955	0.02479575405994061
	0.0007722413981045025	0.07106742542200244
	0.00035744947906382667	0.0977536406001604
	0.0004430660691492 67	0.023227950104680175
	0.0004560582070451528	0.05123031502039276
	0.0007443487952630986	0.06122782435629702

Hierarchical ,word2vec



Hierarchical word2vec	Adjusted rand score	Normalized mutual information
	0.021123165650887364	0.18732625827677468
	0.11995052149650147	0.3869936152405817
	0.11135871755728788	0.4018876921321646
	0.00027371968318355325	0.023006729821726045
	0.000159581350015326	0.048416553739119815
	0.00031540635527447787	0.061584284132000046

در سه ردیف اول هر جدول تعداد خوشه‌های متفاوت و در سه ردیف آخر iteration ها متمایز هستند.

همانطور که از نتایج پیداست ؛ از بین سه ردیف اول مدل با ۴۹ خوشه(تعداد خوشه‌های اصلی داده‌ها) و در سه ردیف آخر مدل با تعداد تکرار بیشینه، بهترین نتایج را دارند.

همچنین بین دو مدل word2vec , tf_idf مدل word2vec نتایج بهتری را نشان می‌دهند.