پروژه پایانی درس هوش محاسباتی:

دیتاستی که در اختیار شما قرار داده شده است مربوط به تعدادی فیلم می باشد که از سایت IMDB دریافت شده است. هر یک از فیلم ها یک خلاصه داستان دارد، وبه یک یا چند ژانر تعلق دارد. دیتاست به دو بخش تست و ترین تقسیم شده است که هر کدام در یک فایل csv مجزا قرار گرفته اند. در این پروژه میخواهیم با استفاده از الگوریتم های پردازش زبان طبیعی که در کلاس بررسی شدند، سیستمی طراحی کنیم که ژانر هر فیلم را با استفاده از خلاصه داستان آن حدس می زند.

1	[{'id': 35, 'name': 'Comedy'}]	When Lou, who has become the "father of the Internet," is shot by an unknown assailant, Jacob and Nick fire up the time machine again to save their friend.
2	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'name': 'Drama'}, {'id': 10751, 'name': 'Family'}, {'id': 10749, 'name': 'Romance'}]	Mia Thermopolis is now a college graduate and on her way to Genovia to take up her duties as princess. Her best friend Lilly also joins her for the summer. Mia continues her 'princess lessons'- riding horses side-saddle, archery, and other royal. But her complicated life is turned upside down once again when she not only learns that she is to take the crown as queen earlier than expected...
3	[{'id': 18, 'name': 'Drama'}]	Under the direction of a ruthless instructor, a talented young drummer begins to pursue perfection at any cost, even his humanity.
4	[{'id': 53, 'name': 'Thriller'}, {'id': 18, 'name': 'Drama'}]	Vidya Bagchi (Vidya Balan) arrives in Kolkata from London to find her missing husband Arnab Bagchi. Seven months pregnant and alone in a festive city, she begins a relentless search for her husband. With nothing to rely on except fragments from her memories about him, all clues seem to reach a dead end when everyone tries to convince Vidya that her husband does not exist. She slowly realises that nothing is what it seems. In a city soaked in lies, Vidya is determined to unravel the truth about her husband - for herself and her unborn child - even at the cost of her own life.

نمونه ای از فیلم های دیتاست

این پروژه سه بخش دارد که به شرح زیر می باشند:

۱. استفاده از word2vec:
از الگوریتم word2vec که از قبل ترین شده است، برای بازنمایی کلمات خلاصه داستان ها استفاده کنید. برای بازنمایی کل پاراگراف از میانگین بازنمایی های کلمات آن استفاده کنید. پس از بدست آوردن این بازنمایی، یک کلسیفایر روی داده های ترین آموزش دهید که ژانر یا ژانر های هر فیلم داده تست را تعیین کند.

۲. BoW:
در قسمت دوم میخواهیم همان تسک قبل را این بار با کمک Bag of Words انجام دهیم. با توجه به مطالبی که در درس خواندید، بازنمایی هر پاراگراف در دیتاست را با استفاده از tf-idf بدست آورید، و یک کلسیفایر روی داده های ترین آموزش دهید. دقت کلسیفایر روی داده های تست در این قسمت را با دقتی که که از بخش قبلی بدست آورید مقایسه کنید و نتایج را تحلیل کنید.

۳. بهبود الگوریتم ها:
روشی برای ترکیب دو الگوریتم قبل پیشنهاد دهید که عملکرد آن بر روی دیتاست از هر یک به تنهایی بهتر باشد. مشابه دو قسمت قبل، الگوریتم پیشنهادی خود را روی داده های ترین آموزش دهید و دقت آن را بر روی دیتا تست بررسی کنید. برای هر سه قسمت کلسیفایر های مختلف را امتحان کرده و عمکرد هر کدام را بررسی کنید.

دقت داشته باشید که اگر فیلمی دو یا چند لیبل دارد، آن را با هر لیبل بصورت مجزا در نظر بگیرید. برای مثال در داده چهارم جدول لیبل فیلم هم Thriller است و هم Drama. آن را با تنها یک لیبل Thriller/Drama نشان ندهید.

برای انجام این پروژه تا پایان روز سه شنبه ۱۵ تیر ماه فرصت دارید. لطفا تا این تاریخ کد و گزارش خود را در سایت ویو آپلود کنید. پروژه به صورت تک نفره می باشد و می توانید از متلب یا پایتون برای پیاده سازی آن استفاده کنید. استفاده از کتابخانه های آماده بلامانع می باشد.