clustering method for the project:

هدف ما انتخاب چندین شغل برای پیشنهاد کردن به کاربر هست به صورتی که بدانیم به احتمال زیاد کاربر به ان شغل ها علاقه نشان میدهد یا برایش مفید میباشد.

هدف کار ما این هست که ما ابتدا میخواهیم تمام یوزر ها و یا کاربر ها را به یک وکتور تبدیل کنیم و برای این کار به چندین مشخصه یا مهارت کاری نیاز داریم برای همین چندین مشخصه یا ویژگی برای کاربر تعریف میکنیم که کاربر در هنگام ثبت نام باید پر کند و از یک تا 10 به خودش در ان مهارت نمره بدهد. این مشخصه ها میشوند ابعاد وکتور ما مثلا چندین مشخصه مثل مهارت برنامه نویسی و فن بیان یا کنترل استرس و .. تعریف میکنیم و یوزر به خودش نمره میدهد و همچین برای تمام شغل ها کاربر باید بتواند که یک نمره از یک تا ده بدهد به این عنوان که این اگهی شغلی چقدر برایش مفید بوده است .

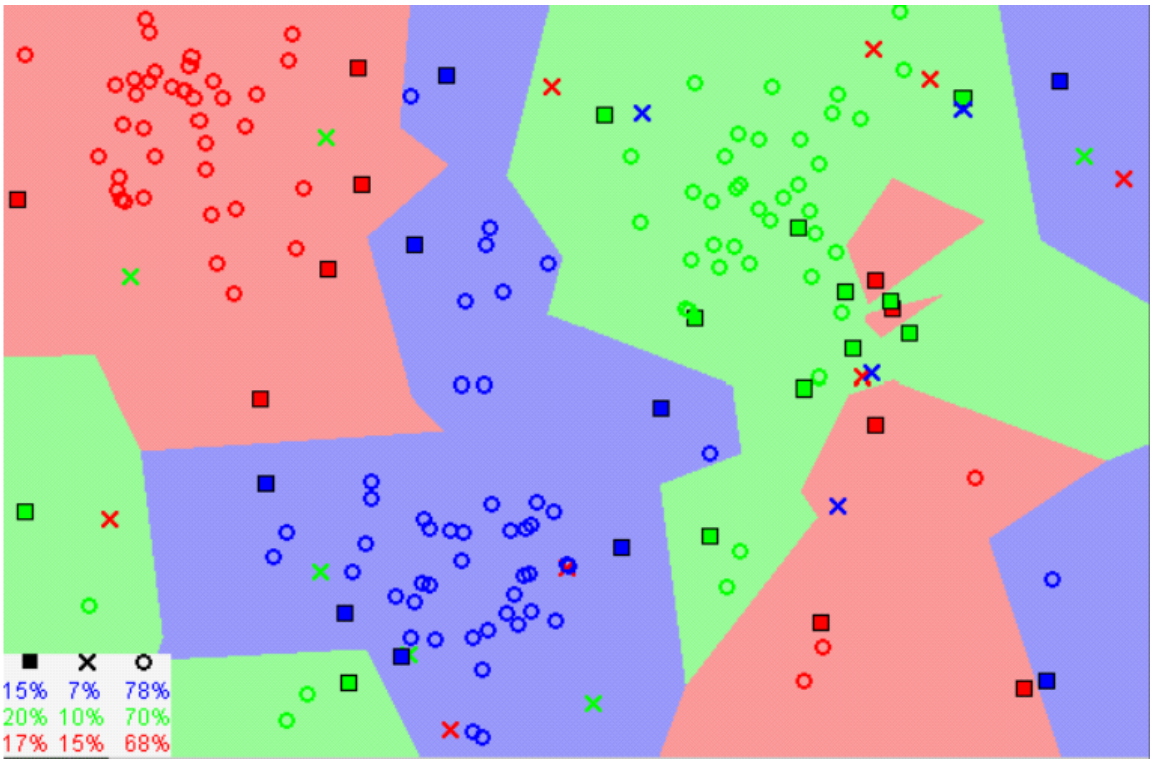
هدف ما از ویژگی های کاربر این است که بتوانیم کاربران شبیه به هم را شناسایی کنیم و هدف ما از فیلد امتیاز هر کاربر به شغل های دیده شده توسط همان کاربر این است که کاربران نزدیک به هم در یک دسته چقدر به مشاغل دیده شده علاقه نشان داده اند یا برایشان این تبلیغ مهم بوده است.

بعد از ان ما تمام یا بخشی از یوزر هارا انتخاب میکنیم یا همان دیتاست خودمان را انتخاب کرده در فضای برداری خودمان و از الگوریتم

k-means clustering

برای کلاستر و دسته بندی کردن انها استفاده میکنیم که اصطلاحا فضای داده های تمرینی ما را به چندین بخش تقسیم میکند.

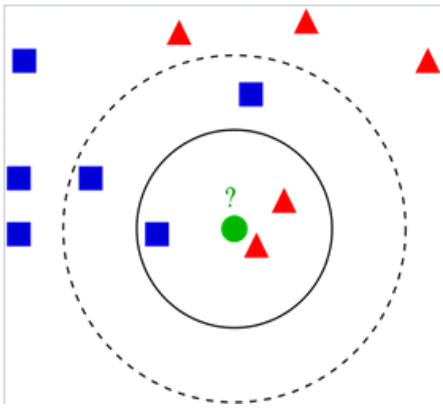
میتوان این یوزر ها را خودمان جنریت یا تولید کنیم و یا به صورت رندوم بخشی از ان هارا انتخاب کنیم.



حالا بعد از انجام دادن این الگوریتم روی دیتاست خودمان هنگامی که یوزر جدید ثبتنام میکند و در هنگام ثبت نام به خودش و ویژگی های خودش یا همان مشخصه های تعیین شده ما نمره میدهد برای شناسایی دسته کاربر از الگوریتم

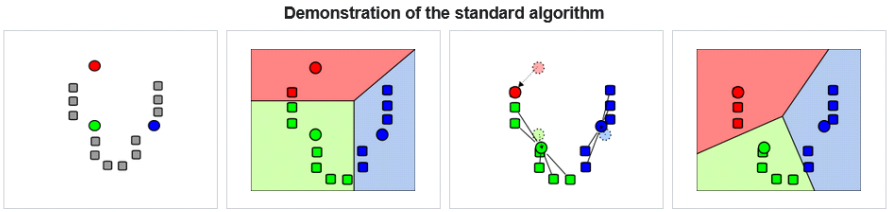
k-nearest neighbors algorithm

استفاده کرد یا میتوان الگوریتم خوشه بندی را دوباره اجرا کرد و دوباره تکرار کرد که پیشنهاد نمیشود چون خیلی از لحاظ عملیاتی هزینه بر میباشد.



روش کار الگوریتم کلاسترینگ ما با این نحو هست که ابتدا چندین نقطه یا یوزر رندوم را انتخاب میکنیم و این گونه برایش محاسبه میکنیم :

به ازای هر نقطه ای در فضای برداری ما ما نزدیک ترین نقطه نقاط رندوم اولیه انتخاب شده را انتخاب میکنیم و بعد از تمام این کار وسط ان کلاستر هارا محاسبه کرده و دوباره این کار را انجام میدهیم تا جایی که تغییری در کار ما انجام نشود.



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

بعد از انکه کار کلاسترینگ انجام شد و از الگوریتم نزدیک ترین همسایگی استفاده کردیم متوجه میشویم که یوزر ما در کدام دسته قرار دارد بعد از ان کاری که میکنیم این است که باید این فرمول را بسنجیم که میزان شباهت کاربر با همدسته ای هاش خودش را میدهد کاری که این فرمول میکند متوجه میشود که کاربر موردنظر ما با کاربر های داخل کلاستر و گروه شبیه خودش چقدر شباهت دارد اینگونه درصد شباهت را لحاظ میکنیم و وزن بیشتری به شبیه ترین کاربرها به یوزر موردنظرمان میدهیم:

$$w_{a,u} = \frac{\sum_{i \in I} (r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u)}{\sqrt{\sum_{i \in I} (r_{a,i} - \bar{r}_a)^2 \sum_{i \in I} (r_{u,i} - \bar{r}_u)^2}} \qquad (1)$$

where I is the set of items rated by both users, $r_{u,i}$ is the rating given to item i by user u, and $\bar{r}_u$ is the mean rating given by user u.

و بعد از ان به ازای هر شغل ما عدد ان شغل را اینگونه به دست می اوریم که :

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u \in K} (r_{u,i} - \bar{r}_u) \times w_{a,u}}{\sum_{u \in K} w_{a,u}} \qquad (2)$$

و بعد از ان جدول اعداد شغل هارا سورت میکنیم و پنج شغل اولی که یوزر در حال حاضر ما ندیده را برایش نشان میدهیم.

نیاز های دیتابیس ما هم اینگونه هست که هر کاربر باید برایش برای مثال بیست مشخصه تعریف شود و هر مشخصه از یک تا ده نمره بگیرد اینگونه شباهت بین هرکاربر را میتوانیم مشخص کنیم همچنین باید در دیتابیس هر کاربر امتیازاتی که به مشغال مختلف داده هست وجود داشته باشد که این هم بین 1 تا 10 هست و در حالتی که امتیاز نداده به صورت پیشفرض 0 میگذاریم اینگونه میتونیم بفهمیم که کاربر های شبیه به کاربر مورد نظر از چه مشاغلی راضی بوده اند و فرمول بالا را در مورد انها به کار میبریم و از فرمول بالا چندین شغل را انتخاب کرده و برای مثال ادرس پنج شغلی که بیشترین امتیاز را داشته اند یعنی نزدیک تر به 10 بوده اند را برمیگردانیم و این کار را میتوانیم هر بار که کاربر به سایت لاگین کرد انجام دهیم چون دسته کاربر را شناسایی کردیم و کاربر به یک سری از مشاغل جدید نمره داده پس فرمول بالا بعد از هربار لاگین کردن کاربر میتواند تکرار بشود و همچنین هر یوزر باید دارای یک فیلد باشد که دسته بندی کلاستر شده اش را نشان بدهد.

منابع:

http://blog.christianperone.com/2011/09/machine-learning-text-feature-extraction-tf-idf-part-i/

https://ethen8181.github.io/machine-learning/clustering/tfidf/tfidf.html#Nearest-Neighbors

https://nbviewer.org/github/rasbt/pattern_classification/blob/master/machine_learning/scikit-learn/tfidf_scikit-learn.ipynb

https://www.inf.uni-hamburg.de/en/inst/ab/lt/publications/2018-elsaftyetal-naacl-industry.pdf

https://towardsdatascience.com/recommender-engine-under-the-hood-7869d5eab072

https://www.itm-conferences.org/articles/itmconf/pdf/2022/04/itmconf_icacc2022_02006.pdf

https://ambarishg.github.io/posts/recommender-career-tfidf/

https://practicaldatascience.co.uk/data-science/how-to-create-content-recommendations-using-tf-idf

https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

https://en.wikipedia.org/wiki/Ball_tree

https://en.wikipedia.org/wiki/Tf%E2%80%93idf

https://en.wikipedia.org/wiki/Recommender_system

https://arxiv.org/ftp/arxiv/papers/1301/1301.7363.pdf

https://cran.r-project.org/web/packages/rrecsys/vignettes/b4_funkSVD.html

https://towardsdatascience.com/matrix-factorization-in-recommender-systems-3d3a18009881

https://www.youtube.com/watch?v=4b5d3muPQmA

https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

https://en.wikipedia.org/wiki/K-means_clustering

https://medium.com/sfu-cspmp/recommendation-systems-user-based-collaborative-filtering-using-n-nearest-neighbors-bf7361dc24e0

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

https://medium.com/fnplus/neighbourhood-based-collaborative-filtering-4b7caedd2d11

https://blog.acolyer.org/2019/02/18/the-why-and-how-of-nonnegative-matrix-factorization/

https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

https://www.researchgate.net/publication/200121027_Collaborative_Filtering_Recommender_Systems

https://web.archive.org/web/20060527214435/http://ectrl.itc.it/home/laboratory/meeting/download/p5-l_herlocker.pdf

https://forem.julialang.org/mroavi/naive-k-means-39dk

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

https://www.prem-melville.com/publications/recommender-systems-eml2010.pdf

Collaborative Filtering in Job

Recommender System

Current Topics of Data Engineering

Under the guidance of Prof. Stefan Kettemann

A Survey of Collaborative Filtering Techniques

Xiaoyuan Su and Taghi M. Khoshgoftaar