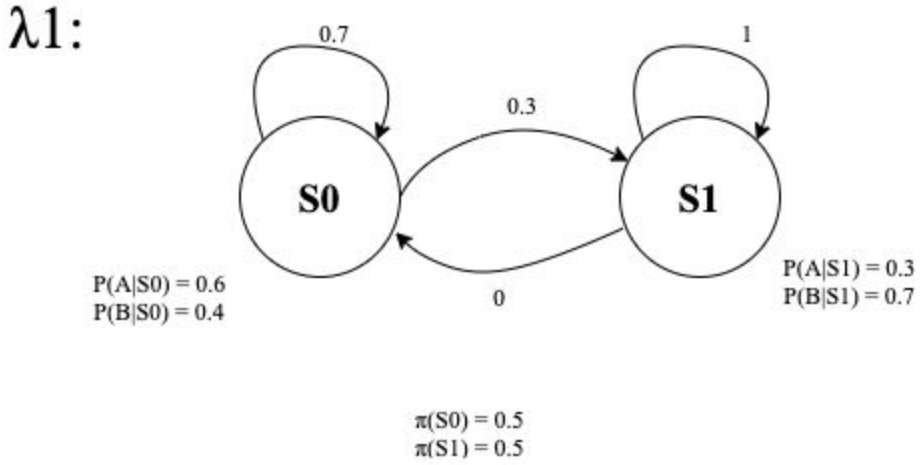
سوال ۱-

پیاده‌سازی تقطیع‌گر انجام شده است؛ البته باید ویژگی‌های بیشتر و بهتری استخراج کرد زیرا هنوز دقت تقطیع‌گر پایین است. بالاخره موفق شدیم یک تجزیه‌گر سازه‌ای پیدا کنیم و قدم بعدی در پیاده‌سازی تقطیع‌گر مطالعه مقاله میگرمن (۱۹۹۵)[1] برای گنجاندن ویژگی‌های نحوی و هم‌چنین افزودن ویژگی‌ها و بهتر کردن آن‌هاست.

برای پیاده‌سازی تجزیه‌گر نیز، مقاله‌ی جی و آیزنشتاین (۲۰۱۴)[2] دوباره مطالعه شد و برای درک مقاله، تا حدی به مطالعه جبر خطی پرداختیم. کد این تجزیه‌گر در دسترس عموم است. ما نیز کد را دریافت کردیم و برای زبان انگلیسی از آن اجرا گرفتیم. حال باید تمام پیش‌نیازهای لازم برای زبان فارسی (در نسخه انگلیسی از پکیج stanfordcorenlp استفاده می‌شود. ما نیز باید هر آن‌چه این پکیج دارد، گرد هم آوریم و به سیستم بدهیم) مهیا شوند و خروجی تقطیع‌گر به تجزیه‌گر داده شود. بنابراین، هزینه‌برترین قدم بعدی آماده‌سازی این پیش‌نیازهاست.

سوال ۲-

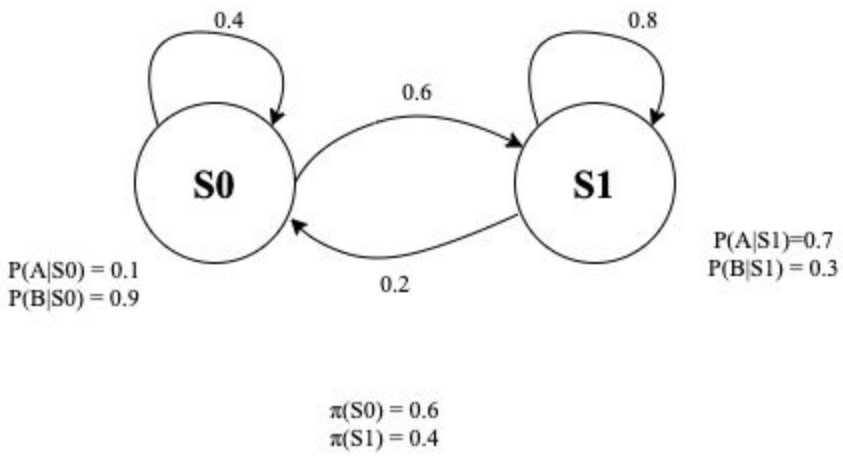
الف)



[1] De Marcken, C. (1995). Lexical heads, phrase structure and the induction of grammar. In *Third Workshop on Very Large Corpora*.

[2] Ji, Y., & Eisenstein, J. (2014). Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 13-24).

λ2:



ب) برای مقادیر اولیه از فرمول $\alpha_0(i) = \pi_i$ استفاده کردیم؛ طبق اسلایدهای درس. یعنی در حالت شروع، emission نداریم.

- **احتمال دنباله ABAB در مدل اول با استفاده از الگوریتم فوروارد:**

مقدار $P(ABAB|\lambda 1)$ را با ترلیس زیر محاسبه کردیم:

	Start	A	B	A	B
S0	0.5	0.21	0.0588	0.024696	0.00691488
S1	0.5	0.195	0.1806	0.059472	0.04681656

پس $P(ABAB|\lambda 1)$ برابر است با جمع ستون آخر یعنی:

$P(ABAB|\lambda 1) = 0.00691488 + 0.04681656 = 0.05373144$

- **احتمال دنباله ABAB در مدل دوم با استفاده از الگوریتم فوروارد:**

مقدار $P(ABAB|\lambda 2)$ را با ترلیس زیر محاسبه کردیم:

	Start	A	B	A	B
S0	0.6	0.032	0.0972	0.006288	0.021708
S1	0.4	0.476	0.12	0.108024	0.0270576

پس $P(ABAB|\lambda 2)$ برابر است با جمع ستون آخر یعنی:

$P(ABAB|\lambda 2) = 0.021708 + 0.0270576 = 0.0487656$

❖ بنابراین، احتمال دنباله ABAB در مدل اول بیشتر است.

ج) محاسبه بهترین دنباله حالت برای ABAB در مدل دوم:

برای مقادیر اولیه از فرمول $\delta_i = \pi_i b_i(o_1)$ استفاده کردیم؛ طبق اسلایدهای درس.

مسیر سبز، مسیر بهینه نهایی است.

فلش‌های قرمز نشان می‌دهد در ماکسیمم‌گیری ویتربی کدام خانه قبلی انتخاب شده است.

	A	B	A	B
S0	0.06	0.0504	0.002016	0.00677376
S1	0.28	0.0672	0.037632	0.00903168

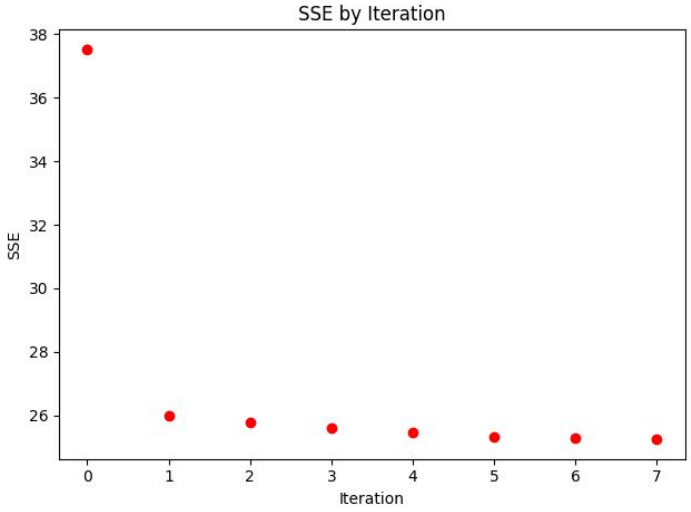
بنابراین، بهترین دنباله حالت برابر است با S1,S1,S1,S1.

سوال ۳-

الف)

برای این تمرین کلاس KmeansClustering تعریف شد که با آدرس دیتاست، تعداد k، تولرانس و حداکثر تکرارهای الگوریتم (maxIteration) مقداردهی اولیه می‌شود و آبجکتی از آن ساخته می‌شود. متدهای پیش‌پردازش داده (مثل تابع _normalize و غیره) و متد selectLexicon از تمرین‌های قبل نوشته شده‌اند، بنابراین توضیح دوباره‌ای درباره‌شان نمی‌دهیم.

- متد doc2vec: این متد بردار TF لگاریتمی سند را می‌سازد.
- متد kmeansClustering: این متد خوشه‌بندی را انجام می‌دهد و نمودار میزان خطا را بر حسب iteration رسم می‌کند. ورودی این متد تمام نمونه‌ها و هم‌چنین یک پارامتر drawErrorPlot -که دو مقدار بولی می‌گیرد- است؛ در صورتی که می‌خواهید نمودار میزان SSE در هر تکرار را در نموداری ببینید، مقدار این پارامتر را True کنید. یک مثال از این نمودار در زیر آمده است:



- تابع computePurity: این متد برای محاسبه خلوص خوشه‌ها نوشته شده است. ورودی آن

ب) برای گزارش کردن میزان خلوص در ۷ خوشه، ۱۴ خوشه و ۷۰ خوشه، برنامه ۱۰۰ بار (با سیدهای اولیه متفاوت) اجرا شد و میانگین خلوص در این ۱۰۰ بار برای هر k در زیر آمده است.

Purity	**k**
0.46	7
0.63	14
1	70

می‌بینیم که خلوص با افزایش k افزایش می‌یابد؛ و وقتی k برابر با ۷۰ (تعداد نمونه‌ها) است، خلوص برابر با ۱ است. اما همان‌طور که از پیش می‌دانیم تعداد کلاس‌ها ۷ است و خوشه بندی با k=70 یا k=14 بهتر از خوشه‌بندی با k=7 نیست؛ این یک ایراد بزرگ معیار خلوص در ارزیابی خوشه‌بندی است. به نظر می‌رسد وقتی تعداد k را از تعداد کلاس‌ها بیشتر می‌گیریم، خلوص نمی‌تواند ایراد خوشه‌بندی ما را تشخیص دهد زیرا تعداد k را در نظر نمی‌گیرد، بنابراین همیشه معیار خوبی برای ارزیابی خوشه‌بندی نیست؛ مثلا، در همین مثال، زمانی که تعداد خوشه‌ها آن‌قدر زیاد شود تا به تعداد نمونه‌ها برسد، خلوص قادر نیست ایراد خوشه‌بندی را ببیند.