



## دانشگاه اصفهان

### درس بازیابی پیشرفته اطلاعات

### فاز ۳ پروژه پایانی

خرداد ۱۴۰۱

زمان تحویل: ۱۴۰۱/۳/۲۳

احتمالا همه شما با موسسه [TED](#) آشنا هستید. همانطور که می‌دانید این وبسایت اقدام به برگزاری رویدادهای متعددی در سراسر جهان می‌کند که در آنها افراد مختلف در حوزه‌های گوناگون مانند آب و هوا، فناوری، زیست، شیمی و غیره از تجارب و پیش‌بینی‌های خود صحبت می‌کنند. همچنین تمامی محتوای تولیدشده توسط این وبسایت به صورت رایگان و از طریق وبسایت آنها قابل دسترسی هستند (چه آدمای خوبی 😊).

در حال حاضر دادگان بسیار متعددی با اهداف مختلف مانند یادگیری ماشین<sup>۱</sup>، پردازش زبان طبیعی<sup>۲</sup>، معنانشناسی<sup>۳</sup> و غیره بر اساس محتوای موجود در [ted.com](#) ایجاد شده‌اند. یکی از این دادگان (شاید ساده‌ترین آنها!) که به فرمت CSV است داده‌های مربوط به ۵۴۴۰ عنوان سخنرانی مختلف به زبان انگلیسی می‌باشد. این دادگان (با حجم 863.62 kb) در وبسایت [kaggle](#) قرار داده شده است. اسکیمای<sup>۴</sup> این دادگان به صورت زیر است:

**title:** The title of the talk

**author:** Author of the talk

**date:** The date when the talk took place

**views:** The number of views of the talk

**likes:** The number of likes of the talk

**link:** The link of the talk in [ted.com](#)

---

<sup>1</sup> Machine Learning

<sup>2</sup> NLP: Natural Language Processing

<sup>3</sup> Semantic Detection

<sup>4</sup> Schema

هدف ما در این پروژه این است که بتوانیم با استفاده از مطالبی که در درس بازیابی پیشرفته اطلاعات آموخته‌ایم، سخنرانی‌های موجود در دادگان اشاره شده را کلاس‌بندی<sup>۵</sup> کنیم. اگر سری به یکی از سخنرانی‌های موجود در [ted.com](https://www.ted.com) بزنید (برای مثال [این سخنرانی](#))، مشاهده خواهید کرد که هر سخنرانی دارای برچسب‌های<sup>۶</sup> مختلفی است که موضوع آن را نشان می‌دهند.



TEDWomen 2021 • December 2021 | 551K views

Like (16K) Share Add

## Climate action needs new frontline leadership

Ozawa Bineshi Albert

We can't rely on those who created climate change to fix it, says climate justice organizer Ozawa Bineshi Albert. An Indigenous woman living in the heart of oil and gas country in the US, she's observed an alarming disconnect between empty promises made by corporations and the actual needs of communities on the ground. In this call for urgency and a shift in values, she advocates for climate policy to center frontline leaders and outlines some grassroots-led projects -- from water protection efforts in Minnesota to off-grid solar power in Arizona -- that have already sparked real change.

[Climate Change](#), [Global Issues](#), [Activism](#), [Fossil Fuels](#), [Indigenous Peoples](#), [Environment](#), [Natural Resources](#), [Policy](#), [Sustainability](#), [Social Change](#)

Read transcript

الصاق این برچسب‌ها به هر یک سخنرانی‌ها، توسط فردی که ویدئو را در [ted.com](https://www.ted.com) آپلود می‌کند، انجام می‌شود. این فرد برای اینکه بتواند این کار را انجام دهد باید چندین بار سخنرانی را گوش دهد و با توجه به برداشتی که از سخنرانی دارد برچسب‌های مناسبی را انتخاب کند. چه فرآیند حوصله‌سربری ☹!! به نظر تان بهتر نیست که بعد از آپلود ویدئو هر سخنرانی، فرآیند الصاق برچسب‌ها به صورت خودکار انجام شود. چه کار سختی مگه نه؟! اما ما در این پروژه به دنبال این موضوع هستیم ☺. همچنین به این نکته توجه داشته باشید که این پروژه آموزش محور است و نه نتیجه محور. یعنی به دنبال یادگیری هستیم تا نتیجه ولی خب نتیجه‌های خوب همیشه شایسته تقدیرند.

### گام اول: شاخص‌گذاری سخنرانی‌ها (Warm-Up)

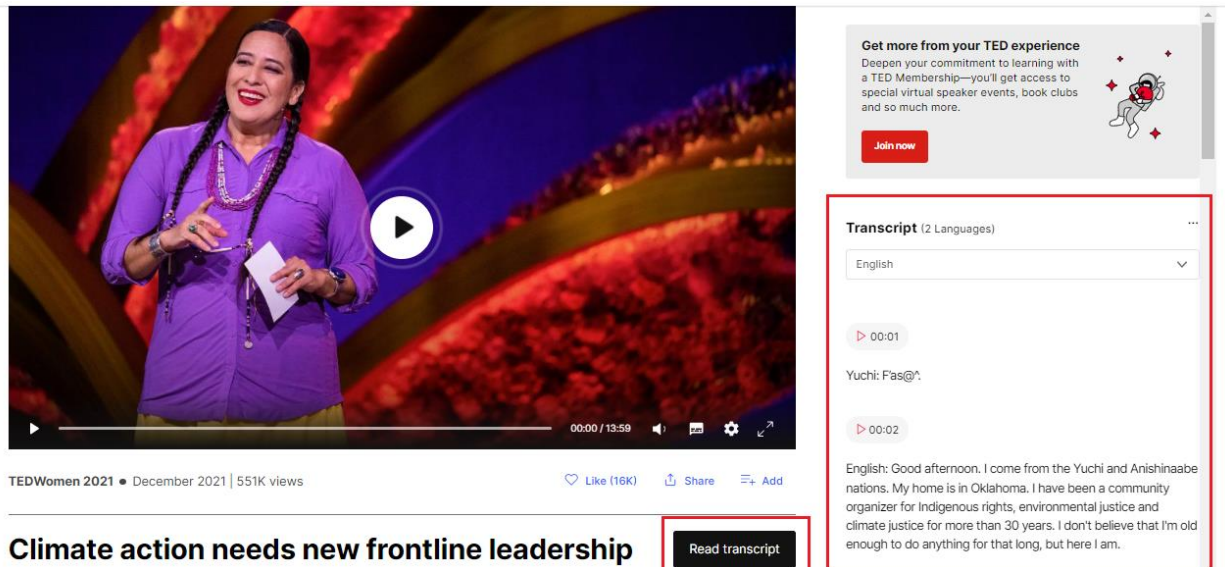
در این گام شما می‌باید هر سخنرانی را با استفاده از Elasticsearch شاخص‌گذاری کنید. Elasticsearch یک موتور جستجوی متن‌باز بر پایه Lucence است. این موتور جستجو یک RESTful web service راه‌اندازی می‌کند که درخواست‌ها به این وب‌سرور ارسال می‌شوند. همچنین قابلیت توزیع گسترده و مقیاس‌پذیری سریع را نیز فراهم می‌کند. برای آشنایی بیشتر با این موتور جستجو، می‌توانید از این [لینک](#) استفاده کنید.

<sup>5</sup> Classification

<sup>6</sup> Labels

کارهایی که باید در این گام انجام شوند:

- ایجاد شاخص talks در Elasticsearch
- شاخص‌گذاری سخنرانی‌های موجود در هر سطر از دادگان یا همان فایل data.csv: شما باید به هنگام شاخص‌گذاری هر یک از سخنرانی‌ها از فیلد link موجود در دادگان استفاده کنید و با نوشتن یک crawler ساده متن سخنرانی را نیز به سند اضافه شده به Elasticsearch به عنوان فیلد transcript اضافه کنید (اگر سری به یکی از سخنرانی‌های موجود در ted.com بزنید مشاهده می‌کنید که متن سخنرانی قابل دسترس است).



The screenshot shows a TED talk video player. The video features a woman in a purple shirt speaking on a stage with a colorful, abstract background. A play button is overlaid on the video. To the right of the video is a sidebar with a 'Join now' button and a 'Transcript (2 Languages)' section. The transcript is in English and shows the beginning of the talk. Below the video, the title 'Climate action needs new frontline leadership' is displayed, along with a 'Read transcript' button.

همچنین به هنگام شاخص‌گذاری متن سخنرانی‌ها باید stop words را حذف کنید که برای اینکار می‌توانید از Analyzerهای Elasticsearch استفاده نمایید.

تحویل‌دانی‌ها:

برنامه‌ایی که فایل data.csv و مقادیر host و port سرور Elasticsearch را بگیرد و شاخص talks را مطابق با آنچه گفته شد، ایجاد نماید.

### گام دوم: تبدیل قسمت چکیده هر سخنرانی به بردار

یکی از ایده‌های نوینی که امروزه در یادگیری ماشین و داده‌کاوی مورد استفاده قرار می‌گیرد، نمایش متون با استفاده از بردار است. بردارها این قابلیت را دارند که بتوانند اطلاعات پنهان موجود در متون مانند شبهات و معنای آنها را آشکار سازند. برای تبدیل متن به بردار الگوریتم‌های زیادی وجود دارد که به صورت کتابخانه در دسترس هستند. یکی از این کتابخانه‌ها، fastText است که توسط شرکت Facebook توسعه داده شده است. این کتابخانه با استفاده از دو الگوریتم Skip-Gram و CBOW اقدام به تبدیل متون به بردار می‌کند. در این گام از پروژه قصد داریم با استفاده از این ابزار و الگوریتم Skip-Gram اقدام به ساخت بردار بخش transcript سخنرانی‌ها نماییم. در واقع در این گام شما باید بخش talk transcript که در Elasticsearch ذخیره کرده‌اید را به

صورت ورودی به این کتابخانه داده، و فایل خروجی که شامل بردار قسمت چیکده است را دریافت نمایید. [برای این منظور می‌توانید از این راهنما استفاده کنید.](#)

توجه داشته باشید که نیازی به پیاده‌سازی الگوریتم Skip-Gram نیست و شما باید از کتابخانه fastText استفاده نمایید.

### گام سوم: کلاسه‌بندی<sup>۷</sup> سخنرانی‌ها بر اساس بردار transcript

در گام دوم با استفاده از کتابخانه fastText اقدام به ساخت بردار مربوط به transcript هر سخنرانی کردیم. در این گام قصد داریم با استفاده از این بردارها، سخنرانی‌ها را در دسته‌های مختلف کلاسه‌بندی کنیم. برای این کار باید مراحل زیر را انجام دهیم:

۱- برچسب‌گذاری<sup>۸</sup> transcript مربوط به هر سخنرانی: برای این کار باید با استفاده از بردار استخراج شده در گام دوم، ۱۰ کلمه که دارای بیشترین فرکانس هستند (بدون در نظر گرفتن stop words) به عنوان برچسب‌های سخنرانی انتخاب شوند.

۲- تقسیم سخنرانی‌ها به دو دسته train و test: از آنجا که برای کلاسه‌بندی قصد داریم از الگوریتم کلاسه‌بندی با نظارت<sup>۹</sup> ارائه شده توسط fastText استفاده کنیم، باید سخنرانی‌ها را به دو دسته train و test تقسیم کنیم. نسبت تقسیم داده‌ها می‌تواند توسط دانشجو انتخاب شود (معمولاً ۹۰ درصد از داده‌ها را برای train و ۱۰ درصد را برای test انتخاب می‌کنند).

۳- انجام فاز train برای سخنرانی‌ها با استفاده از کتابخانه fastText.

۴- انجام فاز test توسط کتابخانه fastText و محاسبه کمیت‌های precision و recall.

**به منظور جزییات بیشتر در مورد مراحل گفته شده در بالا می‌توانید از این [مقاله](#) استفاده کنید.**

توجه داشته باشید که نیازی به پیاده‌سازی الگوریتم کلاسه‌بندی نیست و شما باید از کتابخانه fastText استفاده نمایید.

### **تحويل دادنی‌ها:**

- سورس‌کد<sup>۱۰</sup> برنامه نوشته شده که امکان برچسب‌گذاری transcript سخنرانی‌ها، تقسیم آنها به دو دسته train و test و کلاسه‌بندی آنها با استفاده از کتابخانه fastText را فراهم می‌کند.
- فایل مستندی که در آن روش انجام کار شرح داده شده باشد.

### **نکات:**

- زبان برنامه‌سازی برای پیاده‌سازی این پروژه می‌تواند هر زبانی باشد، اما اکیدا توصیه می‌شود که به دلیل وجود کتابخانه‌های آماده پردازش زبان طبیعی و منابع فراوان برای زبان‌های جاوا و پایتون، شما نیز از یکی از این دو زبان برای پیاده‌سازی این پروژه استفاده کنید.

<sup>7</sup> Classification

<sup>8</sup> Labeling

<sup>9</sup> Supervised Classification

<sup>10</sup> Source code

- این پروژه می‌تواند در قالب تیم‌های دو نفره انجام شود. لذا همه دانشجویان باید در اسرع وقت اقدام به انتخاب هم‌گروهی‌های خود کنند و اطلاعات اعضای هر یک از گروه‌ها، توسط نماینده گروه برای بنده ایمیل شود.
- ارتباط با بنده از طریق آدرس ایمیل [g.elhamesmaeeli@gmail.com](mailto:g.elhamesmaeeli@gmail.com) امکان‌پذیر است.
- هرگونه تبادل نظر و همفکری با سایر گروه‌ها بلامانع است. اما تقلب ممنوع بوده و با گروه متقلب و تقلب‌شونده با کسر کامل نمره پروژه برخورد خواهد شد.