به نام خداوند جان و خرد

کزین برتر اندیشه بر نگذرد

دانشگاه شهید بهشتی

دانشکده ریاضی

پروژه‌ی کارشناسی

غزل رفیعی

استاد راهنما

خانم دکتر بهار فراهانی

آقای دکتر هادی فراهانی

زمستان ۱۴۰۰

دانشگاه شهید بهشتی

دانشکده ریاضی

پروژه‌ی کارشناسی

تحت عنوان:

خودکارسازی فرآیند استخدام در منابع انسانی سازمان‌ها

۱- استاد راهنما: خانم دکتر بهار فراهانی

۲- استاد راهنما: آقای دکتر هادی فراهانی

به نام خدا

نام و نام خانوادگی: غزل رفیعی

عنوان: خودکارسازی فرآیند استخدام در منابع انسانی سازمان‌ها

استاد راهنما: خانم دکتر بهار فراهانی

استاد راهنما: آقای دکتر هادی فراهانی

نام ونام خانوادگی دانشجو: غزل رفیعی

امضاء وتاریخ:

فهرست مطالب

چکیده..... ۱۱
فصل اول: معرفی ۱۲
۱.۱ مقدمه ۱۲
۲.۱ طرح مسأله ۱۲
۳.۱ اهداف ۱۳
۴.۱ سوالات پژوهش ۱۳
۵.۱ محدوده‌ی تحقیق ۱۳
۶.۱ روش و گام‌های پژوهش ۱۴
۷.۱ ساختار گزارش ۱۴
فصل دوم: ادبیات موضوع ۱۵
۱.۲ مقدمه ۱۵
۲.۲ منابع انسانی و فرآیند استخدام ۱۵
۳.۲ سیستم‌های پیشنهادگر ۱۶
۱.۳.۲ مدل بر اساس محتوا ۱۷
۲.۳.۲ مدل بر اساس مشارکت ۱۷
۳.۳.۲ مدل ترکیبی ۱۷
۴.۲ روش‌های تعبیه سازی لغات و متن ۱۸
۵.۲ روش‌های به دست آوردن تشابه بین متون ۲۰
۶.۲ ارزیابی نتایج ۲۱
فصل سوم: پیشینه پژوهش ۲۲

۱.۳ مقدمه ۲۲
۲.۳ دسته‌بندی پژوهش‌های انجام شده در این حوزه ۲۲
۱.۲.۳ پیشنهاد چند رزومه‌ی برتر فرستاده شده برای یک جایگاه شغلی مشخص ۲۳
۲.۲.۳ مرتبط ساختن چند رزومه و چند درخواست ۳۱
۳.۲.۲ یک روزمه و چند شغل ۳۴
۴.۲.۳ یک روزمه و چند شغل ۳۷
۳.۳ جدول مقایسه‌ی کارهای پیشین ۴۰
۴.۳ نتیجه گیری ۴۲
فصل چهارم: رویکرد پیشنهادی ۴۳
۱.۴ مقدمه ۴۳
۲.۴ گام‌های رویکرد پیشنهادی ۴۳
۱.۲.۴ ورودی ۴۴
۲.۲.۴ پیش‌پردازش ۴۵
۳.۲.۴. پیشنهاد دادن ۴۷
فصل پنجم: ارزیابی و اعتبار سنجی ۴۸
۱.۵ مقدمه ۴۸
۲.۵ مورد مطالعاتی ۴۸
۳.۵ نتایج ۴۸
۴.۵ نتیجه‌گیری و جمع‌بندی ۵۰
مراجع ۵۴

فهرست اشکال

شکل ۱.۲- وظایف بخش منابع انسانی به ترتیب ۱۵
شکل ۳.۲- مراحل فرآیند استخدام در منابع انسانی ۱۶
شکل ۴.۲- مقایسه‌ی معماری شبکه‌های Skip-Gram و CBOW [28] ۱۹
شکل ۱.۳- ساختار کلی سیستم مقاله [2] ۲۴
شکل۲.۳- معماری سیستم مقاله‌ی [2] با جزئیات بیشتر ۲۵
شکل ۳.۳- صفحه‌ی پروفایل کاربر[2] ۲۶
شکل ۴.۳- معماری سیسیتم در مقاله‌ی [6] ۲۹
شکل ۵.۳- معماری سیستم در مقاله‌ی [9] ۳۰
شکل ۶.۳- شبکه‌ی عصبی Siamese [13] ۳۳
شکل ۷.۳- معماری سیستم در مقاله‌ی [20] ۳۶
شکل ۸.۳- معماری سیتم مقاله‌ی [26] ۳۹
شکل ۹.۳- قالب اطلاعات استخراج شده توسط مقاله[27] ۴۰
شکل ۱.۴- معماری پیشنهادی سیستم پیشنهاددهنده‌ی رزومه ۴۴

فهرست جداول

جدول ۱- مقایسه‌ی روش‌های مختلف برای رتبه‌بندی کاندیداهای استخدام [14] ۲۷
جدول ۲- ارزیابی رتبه‌بندی کاندیداها با در نظر گرفتن سوابق کاری و بدون آن [14]..... ۲۸
جدول ۳- مقایسه‌ی نتایج برای میانگین شغل‌های جاوا و پایتون در سایت indeed و Resumatcher [20] . ۳۶
جدول ۴- مقایسه‌ی مقاله‌های پیشین ۴۲
جدول ۵- مقایسه‌ی ابزارهای استخراج متن از PDF ۴۵
جدول ۶- مقایسه‌ی ابزارهای ترجمه در پایتون ۴۶
جدول ۶- ارزیابی رتبه‌بندی رزومه‌های فارسی با استفاده از روش‌های تعبیه‌سازی متن ۴۹
جدول ۷- جدول ۶- ارزیابی رتبه‌بندی رزومه‌های انگلیسی با استفاده از روش‌های تعبیه‌سازی متن ۴۹
جدول ۸- مقایسه‌ی مدل TF-IDF برای زبان فارسی و انگلیسی ۵۰

چکیده

سازمان‌ها و شرکت‌های بزرگ به طور معمول تعداد زیادی رزومه برای هرجایگاه شغلی خود دریافت می‌کنند. این رزومه‌ها ممکن است دسته‌بندی شده نباشند یا تعدادی از آن‌ها را نتوان در هیچ دسته‌ای قرار داد. بنابراین، خوانش و مرتب‌سازی این رزومه‌ها برای بخش منابع انسانی شرکت‌ها کاری بسیار زمان‌بر است. چه بسا که نتیجه‌ی نهایی نیز رضایت بخش نباشد. در این پژوهش، در پی تحویل دیجیتال، که قصد دارد فرآیندهای سازمانی را - از جمله فرآیند استخدام در منابع انسانی- هوشمندسازی کند، قصد طراحی یک سیستم پیشنهادگر با استفاده از روش‌های یادگیری ماشین داریم. این سیستم، ابتدا متن را از رزومه‌های دریافتی در فرمت‌های مختلف متن استخراج و تمیز می‌کند. سپس به کمک روش‌های متفاوت یادگیری ماشین و پردازش زبان‌های طبیعی رزومه‌ها را برای هر متن جایگاه‌های شغلی یک سازمان بر اساس شباهت دسته‌بندی می‌کند. این سیستم، ابتدا برای رزومه‌های زبان فارسی و سپس متون ترکیبی فارسی و انگلیسی طراحی شده است.

کلیدواژه‌ها—استخدام الکترونیک، سیستم پیشنهادگر، استخراج اطلاعات، تعبیه‌سازی لغات

فصل اول: معرفی

۱.۱ مقدمه

یکی از مهم‌ترین بخش‌ها در منابع انسانی سازمان‌ها، یافتن استعدادها و انتخاب بهترین افراد برای وظایف مختلف است. در بخش منابع انسانی، پس از این‌که نیازمندی‌های یک سازمان و تعداد افراد مورد نیاز برای هر شغل تعیین می‌شود، آگهی‌های استخدام در وب‌سایت‌ها منتشر می‌شود و تعداد افراد زیادی رزومه‌های خود را ارسال می‌کنند. در این مرحله، سازمان می‌بایست بتواند لیست کوتاهی از رزومه‌ها را برای هر آگهی مشخص کند تا سپس وارد مرحله‌ی مصاحبه و استخدام قطعی شود. خوانش و ارزیابی این رزومه‌ها به دلیل نبود استانداردی برای قالب رزومه کاری بسیار زمان‌بر و دشوار است و در نتیجه‌ی نهایی نیز نمی‌توان مطمئن بود بهترین افراد انتخاب شده‌اند. بنابراین نیازی مبرم به سیستمی خودکار برای انجام این کار در بخش منابع انسانی حس می‌شود. این سیستم می‌بایست پس از استخراج متن از فرمت‌های مختلف فایل، متون را مقایسه و آن‌ها را بر اساس شباهت به هر متن جایگاه شغلی رتبه‌بندی کند.

با توجه به ظرفیت بالای سیستم‌های خودکار کامپیوتری، می‌توان این وظیفه را به آن‌ها سپرد. کامپیوترها می‌توانند حجم زیادی از داده را در مدت زمان کمی پردازش کنند و دقت بیشتری نسبت به انسان‌ها داشته باشند.

این فصل کلیات فعالیت‌های انجام شده در حوزه هوشمند سازی فرآیند استخدام را توضیح می‌دهد. در ادامه در بخش ۱-۲ طرح مسئله ارائه شده است. بخش ۱-۳ اهداف و بخش ۱-۴ سوالات را شرح می‌دهد. محدوده پژوهش در بخش ۱-۵ مشخص شده است. بخش ۱-۶ روش پژوهش را توضیح می‌دهد و در پایان این فصل در بخش ۱-۷ ساختار کلی پایان‌نامه توضیح داده شده است.

۲.۱ طرح مسأله

به دلیل تعداد بالای رزومه‌های دریافتی در یک سازمان و تنوع آن ها از نظر تخصص و دسته‌بندی شغلی امکان مطالعه‌ی دقیق و ارزیابی آن‌ها توسط انسان وجود ندارد؛ افرادی که موظف به این کار هستند می‌بایست وقت زیادی را صرف کنند تا بتوانند اولاً این رزومه‌ها را مطالعه کنند، دوماً اطلاعات آن‌ها را به گونه‌ای خلاصه کنند که امکان مقایسه برایشان فراهم شود و در این فرآیند حتما اطلاعاتی هستند که خواننده به آن‌ها توجهی نمی‌کند یا آن‌ها را فراموش می‌کند. این مشکلات سبب می‌شود نتوان بهترین افراد ممکن را برای یک سازمان استخدام کرد. این مسأله در ادامه منتج به ناکارآمدی کارمندان یک شرکت و نیاز به اخراج و استخدام افراد جدید می‌شود که خود دارای هزینه‌ی مالی و زمانی است. برای جلوگیری از این هدررفت انرژی و زمان و هزینه‌ی مالی، قصد داریم بتوانیم سیستمی طراحی کنیم که به صورت خودکار فرآیند خواندن، پردازش رزومه‌ها را انجام دهد و در نهایت با توجه به متون جایگاه‌های شغلی، رزومه‌ها را بر اساس شباهت رتبه‌بندی کند.

۳.۱ اهداف

در این پژوهش قصد داریم پس از خوانش متون رزومه‌های دریافتی برای یک سازمان، ابتدا تا حد امکان اطلاعات را استخراج و آن را به صورت داده‌ی ساختاریافته در جدول ذخیره کنیم. این جدول شامل اطلاعات شخصی مانند سال تولد، شماره تلفن، ایمیل و اطلاعات کاری و تحصیلی مانند دانشگاه‌ها، مقطع تحصیلی و تجربه‌ی کار در شرکت یا کارخانه است.

در مرحله‌ی دوم با توجه به متون جایگاه‌های شغلی، مناسب‌ترین رزومه‌ها را به صورت یک لیست کوتاه رتبه‌بندی شده، به فرد استخدام‌کننده پیشنهاد می‌دهیم تا او بتواند بهترین تصمیم را برای سازمان بگیرد.

۴.۱ سوالات پژوهش

با توجه به طرح مسئله و موارد ذکر شده سوالات تحقیق پیش‌رو عبارتند از:

۱. چگونه می‌توان از پرونده‌های رزومه با فرمت‌های مختلف، اطلاعات را به صورت درست استخراج و آن‌ها را به صورت ساختاریافته ذخیره کرد؟
۲. روش‌های پردازش متون رزومه و آماده‌سازی آن‌ها برای مرحله‌ی پیشنهاد دادن کدام‌ها هستند؟
۳. روش‌های متفاوت سیستم‌های پیشنهاد دهنده کدام‌ها هستند و چگونه می‌توان برای رتبه‌بندی رزومه‌ها بر اساس متن آگهی شغلی از آن‌ها استفاده کرد؟
۴. و در نهایت چگونه می‌توان بخشی از فرآیند استخدام را به کمک پاسخ سوالات بالا، خودکارسازی کرد؟

۵.۱ محدوده‌ی تحقیق

در این پژوهش ابتدا تنها روی رزومه‌های با زبان فارسی کار می‌کنیم. سپس، به علت اینکه در دنیای کاربرد عموماً رزومه‌ها با هر دو زبان فارسی و انگلیسی، یا تنها انگلیسی هستند، سیستم را به گونه‌ای گسترش می‌دهیم که برای هر دو زبان کارآمد باشد. در مورد موضوع رزومه‌ها، از آن جا که در برخی شرکت‌ها و سازمان‌ها رزومه‌ها از همان ابتدا به صورت دسته‌بندی شده بر اساس موضوع نیستند، رزومه‌هایی در چند موضوع مختلف حسابداری، کامپیوتر و نرم‌افزار و علوم داده را بررسی می‌کنیم. بدیهی‌ست که این روش در مورد رزومه‌هایی هم که تنها در یک دسته‌ی موضوعی قرار دارند، کارا خواهد بود.

۶.۱ روش و گام‌های پژوهش

روش تحقیق بر مبنای هدف تحقیق در دسته تحقیق‌های کاربردی قرار می‌گیرد. که با هدف توسعه دانش کاربردی در زمینه سیتم‌های پیشنهاددهنده است. رویکرد ارائه شده در این گزارش شیوه‌ای ابتکاری در روش تحقیق شامل مراحل زیر است:

- تعریف مسئله: تعریف مسئله نقطه آغاز تحقیق و ورود فضای مسئله است و شامل اهمیت مسئله، محدوده تحقیق و سوال‌های تحقیق می‌شود. در تعریف مسئله کلیات و پیش‌زمینه موضوع تحقیق تشریح شده‌است.
- بررسی مفاهیم: آشنایی با مفاهیم و مقدمات لازم برای تحقیق را در بر می‌گیرد. بررسی مفاهیم در فضای مسئله قرار دارد و شامل آشنایی با ادبیات موضوع و جمع‌آوری دانش اولیه برای درک حوزه تحقیق است.
- مقایسه و تحلیل: دانش جمع‌آوری شده برای موضوع تحقیق برای ورود به فضای جواب تحلیل و بررسی می‌شود. مقایسه و تحلیل شامل بررسی و تحلیل حوزه تحقیق و کارهای مرتبط می‌شود و پیوند دهنده فضای مسئله و فضای جواب است.
- استنتاج و پیشنهاد: طرح پیشنهاد برای پاسخ به مسئله تحقیق را در بر می‌گیرد. استنتاج و پیشنهاد در فضای راه حل قرار دارد. ایده و رویکرد جدید برای پاسخ به مسئله تحقیق در این مرحله ایجاد شده است.
- اعتبارسنجی و مقایسه: استنتاج صورت گرفته و پیشنهاد ارائه شده اعتبارسنجی می‌شود. در این بخش از فضای راه حل برای اطمینان از صحت استنتاج صورت و کاربرد رویکرد پیشنهاد شده از طریق روش‌های اعتبارسنجی اعتبار رویکرد پیشنهاد شده بررسی شده است.
- ثبت دستاوردها و جمع‌بندی: دستاوردها، محدودیت‌ها و نتایج حاصل از تحقیق صورت گرفته جمع‌بندی می‌شود. در پایان راه تحقیق تمام جوانب و نتایج تحقیق صورت پذیرفته برای انتشار مهیا و ثبت شده است

۷.۱ ساختار گزارش

در ادامه در فصل دو مفاهیم و ادبیات موضوع تشریح شده و سپس در فصل سوم کارهای پیشین مرتبط با موضوع مورد تحقیق در پایان‌نامه بررسی و تحلیل شده است. فصل چهارم به تشریح و ارائه تفصیلی رویکرد پیشنهادی پایان‌نامه در پنج گام برای جمع‌آوری، یکپارچه‌سازی، پیش‌پردازش، پردازش داده و ارائه پرداخته است و معماری پشتیبان رویکرد پیشنهادی را تشریح می‌کند. فصل پنجم به ارزیابی و بررسی رویکرد ارائه شده از جنبه‌های متفاوت می‌پردازد. فصل ششم به جمع‌بندی و نتیجه‌گیری از تحقیق صورت گرفته در پایان‌نامه و همچنین معرفی کارهای آتی می‌پردازد. در نهایت منابع قرار گرفته است.

فصل دوم: ادبیات موضوع

۱.۲ مقدمه

در این فصل به مقدمات و پیش‌نیاز مطالبی می‌پردازیم که در فصول ۳ تا ۶ از آن‌ها استفاده می‌شود. در بخش ۲.۲ به مطالبی درباره‌ی منابع انسانی و فرآیند استخدام پرداخته می‌شود. در بخش ۳.۲ درباره‌ی سیستم‌های پیشنهادگر توضیحاتی داده می‌شود و به ترتیب در سه بخش بعدی به روش‌های تعبیه‌سازی لغات و متن، روش‌های به دست آوردن تشابه و فرمول‌های ارزیابی نتایج می‌پردازیم.

۲.۲ منابع انسانی و فرآیند استخدام

منابع انسانی افرادی هستند که در شغل‌های یک سازمان کار می‌کنند تا محصولات یا خدمات آن سازمان را آماده و تولید کنند. جذب همکاران جدید، آموزش همکاران، حضور و غیاب، حقوق و پاداش دادن به آنها، ارزیابی عملکر کارکنان توانمند سازی آن‌ها و بستن قراردادها و به طور خلاصه ایجاد محیطی سالم و منصفانه برای کارکنان وظیفه‌ی این افراد است.

شکل ۱.۲ وظایف بخش منابع انسانی به ترتیب

از آن جا که در این پروژه قصد داریم تا بخش استخدام را هوشمندسازی کنیم، به مراحل مختلف این بخش می‌پردازیم.

شکل ۳.۲ مراحل فرآیند استخدام در منابع انسانی

در فرآیند استخدام به ترتیب مراحل شکل۳.۲ را طی می‌کنیم. ابتدا کمبودها و نیازهای شرکت برای جایگزین کردن به جای کارمندان قبلی یا اضافه کردن ظرفیت شناخته می‌شود. برای افراد مورد نیاز توانایی‌های لازم نوشته شده و در قالب متون فرصت‌های شغلی این متون در روزنامه یا شبکه‌های مجازی قرار می‌گیرد تا افراد مرتبط بتوانند برای استخدام اقدام کنند. پس از جمع‌آوری رزومه‌های این افراد، این رزومه‌ها خوانده شده و لیستی کوتاه از میان آن‌ها برای مصاحبه انتخاب می‌شود. در نهایت پس از انجام مصاحبه، در صورت پذیرش افراد، قرارداد به همراه شرایط کار به آن‌ها داده می‌شود و در صورت پذیرش آن‌ها، افراد در شرکت استخدام می‌شوند.

۳.۲ سیستم‌های پیشنهادگر[۱]

امروزه همه‌ی افراد به طریقی با این سیستم ها آشنا هستند. از وبسایت‌های خرید و فروش کالا گرفته تا وبسایت تماشای فیلم یا شبکه‌های اجتماعی. به محض آن که شما در این وبسایت‌ها ثبت نام کرده و شروع به فعالیت می‌کنید، این وبسایت‌های سعی می‌کنند مطالبی مشابه اطلاعاتی که شما تاکنون مشاهده کرده‌اید، به شما نمایش دهند، یا گاهی نمونه‌های جدیدی را به شما پیشنهاد دهند تا سلیقه‌ی شما را بهتر بشناسند و در آینده در زمان شما برای جست و جو در میان مطالب بی‌شمار وبسایت صرفه‌جویی کنند. این سیستم‌ها، سیستم‌های پیشنهادگر نام‌دارند. در این فصل به انواع آن‌ها می‌پردازیم.

---

[۱] Recommender Systems

۱.۳.۲ مدل بر اساس محتوا[1]

این نوع سیستم‌ها تنها تاریخچه‌ی سلیقه‌ی کاربر را در نظر گرفته، و سعی می‌کنند مطالبی مشابه آن‌ها به کاربر پیشنهاد دهند. گرچه پیاده‌سازی این مدل از مدل‌های دیگر ساده‌تر است، اما دو ایراد اساسی دارد.

۱- مشکل اساسی این نوع پیشنهادگرها، مشکل شروع سرد[2] است. به این معنا که هنگامی که کاربر برای اولین بار وارد سیستم می‌شود و هیچ تاریخچه‌ای ندارد، وبسایت نمی‌داند به او چه محتوایی را پیشنهاد کند. یکی از راه‌های حل این مشکل، پرسش از او درباره‌ی سلیقه‌اش قبل از آغاز تجربه‌ی کاربری است.

۲- ممکن است کاربر سلایق دیگری هم داشته باشد اما به دلیل پیشنهاد بیش از اندازه از یک نوع محتوا، مطالب برایش تکراری شود و از وبسایت خارج شود.

۲.۳.۲ مدل بر اساس مشارکت[3]

در این نوع سیستم‌ها، سعی می‌شود بین افراد مختلف بر اساس سلایقشان شباهت پیدا شود و به طور مثال اگر دو کاربر به ترتیب محتواها الف و ب و دیگری الف و پ را دوست داشته باشد، این سیستم محتوای ب را به کاربر دوم و محتوای پ را به کاربر اول پیشنهاد می‌دهد. به این ترتیب مشکل تکراری شدن مطالب برای یک کاربر حل می‌شود.

مشکل شروع سرد در این نوع سیستم‌ها هم وجود دارد.

۳.۳.۲ مدل ترکیبی[4]

در مدل ترکیبی همان طور که از نام آن پیداست، از ترکیب روش‌های محتوامحور و مشارکت‌محور استفاده می‌شود. کارایی این نوع سیستم‌ها بهتر از دو مدل دیگر است و نکته‌ی قابل توجه در این جا این است که با گذر زمان و شناخت کاربران، می‌توان بین بخش های مشارکتی و محتوایی وزن انتخاب کرد تا کاربر بهترین تجربه را از این سرویس داشته باشد.

---

[1] Content-based Recommender System
[2] Cold Start
[3] Collaborative Recommender System
[4] Hybrid Recommender System

۴.۲ روش‌های تعبیه سازی لغات و متن[1]

درون سازی یا تعبیه کلمه نامی تجمعی است که به مجموعه ای از تکنیک های یادگیری ویژگی و مدلسازی زبان در پردازش زبان طبیعی اطلاق میشود. در این تکنیک ها، کلمات و عبارات از یک لغت نامه به بردارهای عددی نگاشت میشوند. بطور مفهومی این تکنیک، مستلزم تعبیه سازی ریاضی از فضایی با ابعاد زیاد به ازای هر کلمه به فضای برداری پیوسته با ابعاد بسیار کمتر است. روشهایی که جهت تولید این نگاشت مورد استفاده قرار میگیرند شامل شبکه های عصبی، کاهش ابعاد بر روی ماتریس هم رخداد کلمه،مدل های احتمالاتی، روش پایه دانش قابل توضیح، و باز نمایی صریح تحت عنوان محتوایی که کلمات در آن ظاهر میشوند، میشود. زمانی که تعبیه کلمه و عبارت، بعنوان بازنمایی ورودی زیرین مورد استفاده قرار گیرد، نشان داده است که سبب افزایش کارایی در کاربردهای مبتنی بر پردازش زبان طبیعی می‌گردد.

دو روش اصلی در رابطه با یادگیری تعبیه‌سازی لغات وجود دارد که هر دوی آنها وابسته به دانش محتوایی اند:

- مبتنی بر شمارش: این روش بدون ناظر[2] بوده و مبتنی برتجزیه ماتریس یک ماتریس هم رخدادی کلمه سراسری است. شمارش هم رخدادی خام به تنهایی بخوبی عمل نمی‌کند و به همین دلیل نیازمند انجام روش هوشمند دیگری بر روی نتایج آن است.

  یکی از روش‌های این دسته TF-IDF نام دارد که مخفف دوم کلمه‌ی Term Frequency و Inverse Document Frequency به ترتیب به معنای تعداد تکرار کلمه در متن و برعکس تعداد تکرار در متون است. این فرمول از روش زیر محاسبه می‌شود.

$$w_{i.j} = tf_{i.j} \times \log\frac{N}{df_i}$$

$tf_{i.j}: number\ of\ occurence\ of\ i\ in\ j$

$df_i: number\ of\ documents\ containing\ i$

$N: total\ number\ of\ documents$

در اینجا i بیان‌گر کلمات و j بیان‌گر متون است.

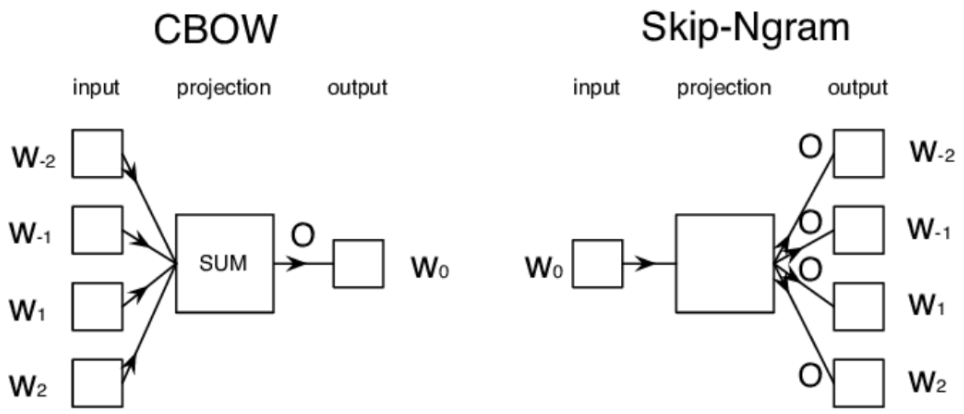
---

[1] Word and Text Embedding

[2] Unsupervised

- مبتنی بر محتوا: این یک روش با ناظر[1] است. در این روش به ازای یک محتوای محلی داده شده، مدلی جهت پیش بینی کلمات هدف طراحی شده و در همین حین، این مدل بازنمایی تعبیه‌سازی لغات بهینه را فرا می‌گیرد.

دو مدل CBOW[2] و Skip-Gram از روش‌های پراستفاده در این دسته هستند. هر دو این مدل‌ها از پنجره‌ای با طول ثابت استفاده می‌کنند و آن را روی متن حرکت می‌دهند و توالی کلمات را می‌آموزند.



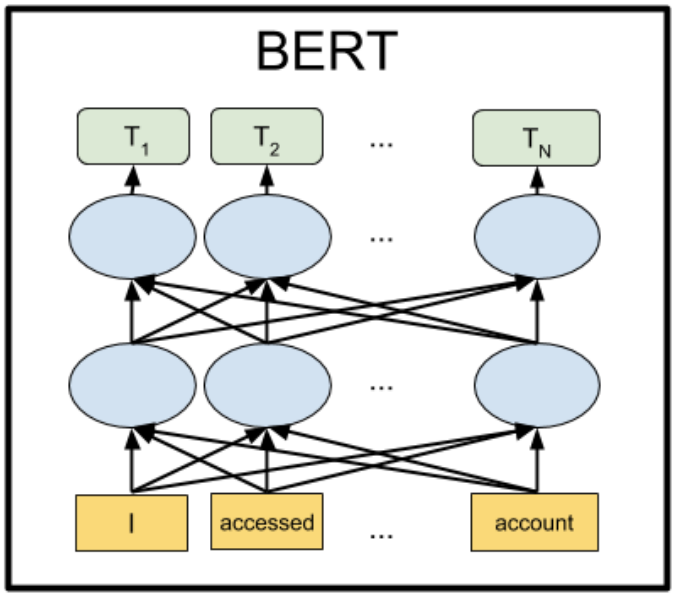
شکل ۴.۲ مقایسه‌ی معماری شبکه‌های Skip-Gram و CBOW [28]

در شکل ۴.۲ می‌توانید معماری این دو شبکه را ببینید. از لحاظ الگوریتمی این دو روش شبیه هم هستند با این تفاوت که CBOW لغات هدف را از روی لغات متن ورودی پیش‌بینی می‌کند ولی اسکیپ‌گرام به صورت برعکس از روی لغات مرجوعه هدف، لغات ورودی را پیش‌بینی می‌کند. برعکس کردن این چرخه دل‌خواه به نظر می‌رسد ولی از لحاظ آماری CBOW تأثیر نرمی بر روی همه اطلاعات توزیعی دارد (با رفتاری شبیه به یک مشاهده بر روی کل متن) و درکل این روش می‌تواند روشی مفید برای استفاده در مجموعه دادگان کوچک‌تر باشد. اما اسکیپ‌گرام با هر زوج محتوا-هدف به‌صورت یک مشاهده جدید رفتار می‌کند و در مجموعه دادگان بزرگ‌تر بهتر جواب می‌دهد.

یکی از روش‌های استفاده از این مدل‌ها، استفاده از ابزار Fasttext[3] ساخته‌ی کمپانی فیسبوک است.

---

[1] Supervised
[2] Continuous Bag of Words
[3] https://fasttext.cc/



شکل ۵.۲- معماری کلی شبکه‌ی عصبی BERT [1]

یکی دیگر از مدل‌هایی که در رابطه با تبدیل متن به برداری از اعداد استفاده می‌شود، شبکه‌ی عصبی BERT است. طراحی این شبکه در شکل ۵.۲ نمایش داده شده است. این شبکه، به جای آنکه پنجره‌ی لغات را تنها از یک طرف حرکت دهد، یک بار از ابتدا به انتها و بار دیگر برعکس حرکت می‌دهد. این عمل باعث می‌شود لغاتی که در ابتدای جملات در متن آمده‌اند، معنای خود را در انتهای جمله از دست ندهند.

۵.۲ روش‌های به دست آوردن تشابه بین متون

پس از آنکه متون به بردارهایی از اعداد تبدیل شدند، مشابهت بین آن‌ها می‌بایست محاسبه شود. روش‌های این کار به همراه فرمول‌هایشان در ادامه توضیح داده شده اند.

- مشابهت کسینوس

$$Cosine\ Similarity(A,B) = \frac{A.B}{|A| \times |B|}$$

- فاصله‌ی اقلیدسی

$$Euclidean\ Distance\ (A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

[1] https://sudonull.com/post/3241-BERT-is-a-state-of-the-art-language-model-for-104-languages-BERT-launching-tutorial-locally-and-on-G

این فاصله درواقع حالت خاصی از فاصله‌ی مینکوفسکی است که به روش زیر حساب می‌شود.

$$Minkowski\ Distance\ (A,B) = \sqrt[p]{\sum_{i=1}^{n}(a_i - b_i)^p}$$

۶.۲ ارزیابی نتایج

برای ارزیابی نتایج در باره‌ی داده‌ی بدون ناظر، روش‌های مختلفی وجود دارد. دو روش معروف در این زمینه عبارتند از دقت kتایی (صحت) و پوشش. روش محاسبه‌ی این دو فرمول به ترتیب در زیر ذکر شده است. در این فرمول‌ها فرض شده از میان تعدادی زیادی گزینه تعدادی از آن‌ها به ما پیشنهاد داده شده و حال قصد داریم میزان موفقیت سیستم برای پیشنهاددهی را بسنجیم.

$$Precision@K = \frac{\text{تعداد موارد مرتبط در } k\text{تای اول}}{\text{تعداد کل موارد پیشنهاد شده}}$$

$$Recall = \frac{\text{تعداد موارد مرتبط در } k\text{تای اول}}{\text{تعداد موارد مشابه واقعی}}$$

هم‌چنین روش سومی وجود دارد که میزان خوب بودن پیشنهادها را در k تای اول می‌سنجد. نحوه‌ی محاسبه‌ی این فرمول در زیر آمده است.

$$Mean\ Avergae\ Precision\ @\ K = \frac{\text{جمع نمرات پیشنهاد شده در } k\text{تای اول}}{\text{جمع نمره کامل برای } k\text{تای اول}}$$

فصل سوم: پیشینه پژوهش

۱.۳ مقدمه

پژوهش‌های انجام شده در این حوزه از سال ۲۰۰۰، همزمان با آغاز همه‌گیری الگوریتم‌های یادگیری ماشین و پیشرفت صنعت کامپیوتر شروع شدند. در آن زمان، بیشتر کارهای انجام شده سعی بر استفاده از الگوریتم‌های ساده‌تر ماتریسی داشتند که عمدتا به صورت غیر موازی انجام می‌شدند و تنها روی حجم محدودی از داده کارایی داشتند. از آن زمان، با پیشرفت و افزایش سرعت کامپیوترها و هم‌چنین پیشرفت در حوزه‌ی موازی‌سازی الگوریتم‌های یادگیری ماشین، روش‌های پیچیده‌تری مانند شبکه‌های عصبی، با پیچیدگی بیشتری قابل پیاده‌سازی روی کامپیوترها شدند و امکان پردازش دقیق‌تر و در حجم بالاتر را برای ما امکان‌پذیر ساختند. این روش‌ها اصطلاحاً هوشمند هستند و می‌توانند در طی زمان با اضافه شدن داده‌های جدید درک خود را از فضای مسأله بهبود ببخشند و خودشان باعث پیشرفت خود بشوند؛ در حالی که در روش‌های قدیمی‌تر ماتریسی، این امکان وجود ندارد و با هربار اضافه شدن داده، می‌بایست الگوریتم روی کل دادگانی که در اختیار داریم، اجرا شود و واضح است که این کار از نظر زمانی به صرفه نیست.

۲.۳ دسته‌بندی پژوهش‌های انجام شده در این حوزه

اگر بخواهیم از دیدگاه منابع انسانی و معماری برنامه به مسأله‌ی هوشمندسازی سیستم استخدام نگاه کنیم، می‌توانیم کارهای پیشین را در چهار دسته جای دهیم. این چهاردسته در بخش تحلیل و پردازش دادگان مشابه هستند و تنها در نحوه‌ی ارائه متفاوت هستند. به همین سبب، در ادامه هر چهار دسته را توضیح می‌دهیم.

۱. پیشنهاد چند رزومه‌ی برتر فرستاده شده برای یک جایگاه شغلی مشخص
۲. مرتبط ساختن چند رزومه و چند درخواست
۳. پیشنهاد شغل برای یک شخص با رزومه‌ی مشخص
۴. مقاله‌های دیگر

در بخش اول، که مشابه پژوهش کنونی ماست، بیشتر در سازمان‌ها و بخش منابع انسانی آن‌ها کاربرد دارد. سیستم‌های در این دسته، برای یک آگهی شغلی مشخص، مرتبط‌ترین رزومه‌ها را پیشنهاد می‌دهند. در بخش دوم، که در وبسایت‌هایی مانند جابینجا[1] و indeed.com به عنوان نمونه‌ی خارجی کاربرد دارند، موظف اند پیشنهاد دو طرفه انجام دهند؛ به این معنا که در این وبسایت‌ها هم متقاضیان شغل و هم متقاضیان کارمند ثبت نام

[1] Jobinja.ir

می‌کنند. این وبسایت‌ها به عنوان شخص ثالث سعی می‌کنند شغل‌ها را به کارمندانی که بیشترین علاقمندی را به آن شغل‌ها، و کارمندان را به کارفرمایانی که بیشترین نیاز را به آن کارمندان دارند پیشنهاد دهند. پیچیدگی این سیستم‌ها از دیگر سه دسته‌ی دیگر کمی بیشتر است زیرا نیاز است که رضایت‌مندی هر دو طرف به بیشترین میزان ممکن فراهم شود.

در بخش سوم، تنها به افرادی که متقاضی شغل هستند مرتبط‌ترین شغل ها با توجه به سابقه‌ی های کاری آن‌ها پیشنهاد داده می‌شود و در بخش چهارم مقالاتی را داریم که کارهای جزئی‌تر مانند تمرکز بر استخراج داده از متن رزومه، یا پیشنهاد شغل بر اساس موقعیت جغرافیایی را انجام داده‌اند.

در ادامه به بررسی مقالات موجود در هرکدام از این بخش‌ها می‌پردازیم و در انتها همه را در یک جدول به صورت خلاصه ارائه می‌دهیم.

۱.۲.۳ پیشنهاد چند رزومه‌ی برتر فرستاده شده برای یک جایگاه شغلی مشخص

- (۲۰۱۹) راهکار یادگیری ماشین برای سیستم پیشنهادگر رزومه [1]

در این مقاله ابتدا روش‌های تمیزکردن داده و استخراج ویژگی با استفاده از NLTK و TF-IDF توضیح داده شده. برای کلاس‌بندی رزومه ها از روش‌های Random Forest، Naive Bayes، Logistic Regression، SVM استفاده شده که بهترین آن‌ها مربوط به SVM است. سپس برای سیستم پیشنهادگر روش‌های Cosine Similarity و KNN آزمایش شده اند.

کتابخانه‌ی Textract[1] برای خوانش فرمت فایل‌های PDF و DOC و برای خلاصه سازی متون کتابخانه‌ی Gensim پیشنهاد شده است. در مورد Gensim[2] گفته شده مقدار ازدست‌رفت زیاد است.

- (۲۰۲۰) چگونه یک پیشنهاددهنده‌ی رزومه بسازیم؟[3]

[1] https://textract.readthedocs.io/en/stable

[2] https://pypi.org/project/gensim

[3] https://towardsdatascience.com/resume-scre)ening-tool-resume-recommendation-engine-in-a-nutshell-53fcf6e6559b

پردازش زبان طبیعی با استفاده از NLTK[1] و SpaCy[2] انجام شده است. برای پارامتری کردن متون و لغات باز هم از TF-IDF استفاده شده است. در این مقاله هم برای سیستم پیشنهادگر از Cosine Similarity استفاده شده است.

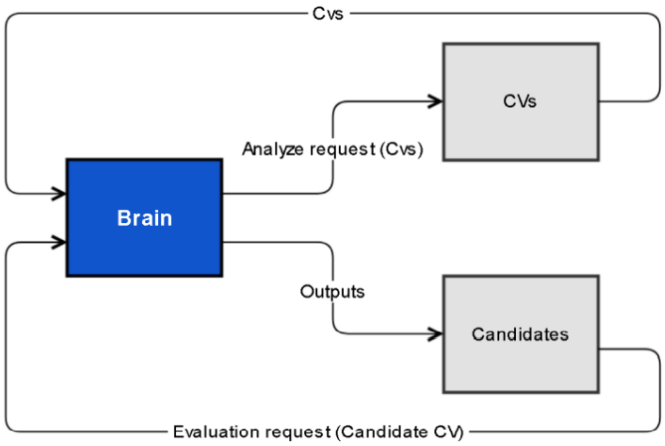
SpaCy و NLTK هر دو از معروف‌ترین کتابخانه‌های پردازش زبان‌های طبیعی هستند. سرعت پردازش SpaCy بسیار بیشتر است.

- (۲۰۲۰) یک پیاده سازی ساده از پروژه[3]

در این صفحه یک پیاده‌سازی از این پروژه را می‌توانیم ببینیم. روش‌های استفاده شده، همان روش‌های گفته شده در مقالات قبلی هستند. داده‌ی این پروژه هم روی سایت Kaggle.com موجود است و در لینک ارجاع داده شده است.

- (۲۰۱۹) پیاده‌سازی ماشین تصمیم‌گیری برای استخدام‌کنندگان[2]

این مقاله از KMeans برای دسته‌بندی رزومه‌ها و از KNN برای یافتن نزدیکترین رزومه‌ها به جایگاه‌های شغلی استفاده می‌کند.



شکل ۱.۳ - ساختار کلی سیستم مقاله [2]

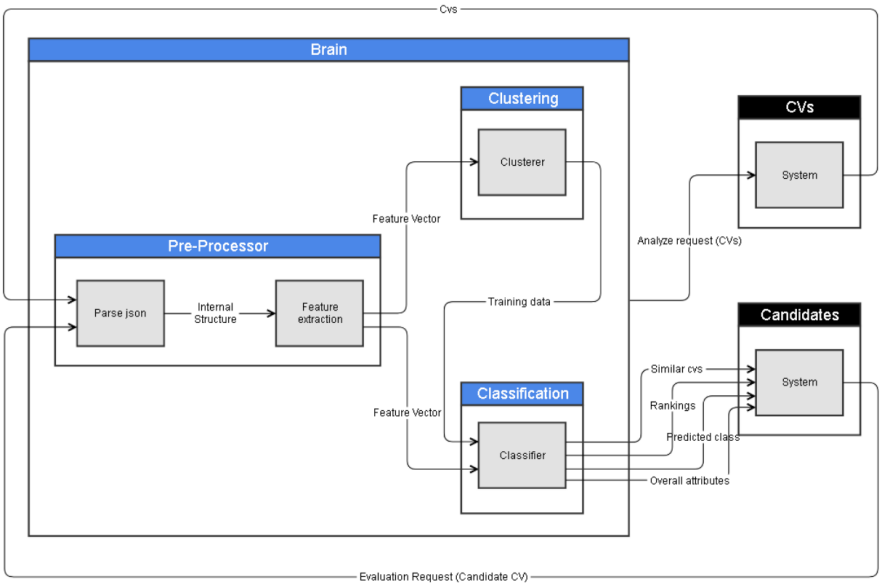
[1] https://pypi.org/project/nltk/

[2] https://pypi.org/project/spacy/

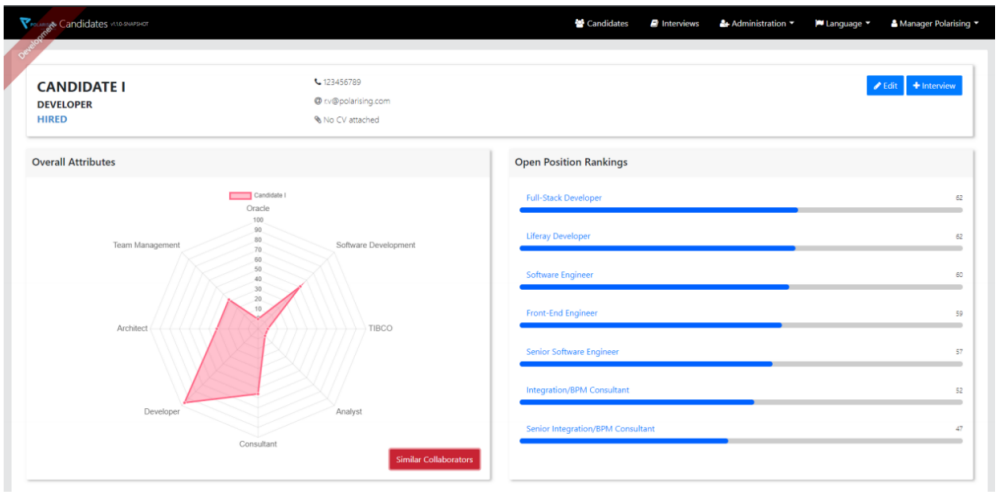
[3] https://github.com/Shailja-Jindal/Bidirectional-Job-Resume-Recommender-System

سیستم طراحی شده دارای معماری مطابق شکل ۱.۳ است که به طور جزئی تر در شکل ۲.۳ بخش‌های مختلف آن نمایش داده شده است. در بخش CVs اشخاص رزومه‌های خود را آپلود و آپدیت می‌کنند. در Candidates خروجی بخش Brain که مرکز پردازش و محاسبات است نشان داده می‌شود. خروجی شامل اطلاعات زیر است:

- ویژگی‌های کلی رزومه‌ها در قالب خاص
- رزومه‌های مشابه
- ترتیب رزومه‌ها از برترین به بدترین در موارد متفاوت مانند سابقه کاری یا تنوع استعدادها یا ...
- پیش‌بینی رزومه‌های خوب برای جایگاه‌ها



شکل۲.۳- معماری سیستم مقاله‌ی [2] با جزئیات بیشتر



شکل ۳.۳- صفحه‌ی پروفایل کاربر[2]

خروجی بخش CV صفحه‌ای مشابه شکل ۳.۳ است. ابتدا اطلاعات شخصی فرد نمایش داده شده، و در پایین می‌توانیم ببینیم کدام توانایی‌ها در این شخص بیشتر هستند.

در ادامه در مورد تخصصی پیاده‌سازی نرم‌افزار توضیح داده شده است. و در انتها برای محاسبه‌ی میزان درستی از Elbow method برای KMeans و Silhouette analysis که مقادیری متناسب با جمع مربعات فواصل از مرکز دسته‌هاست نشان می‌دهد. در انتها با نمایش نمودارهایی گفته شده لزوما اطلاعات بیشتر نتیجه‌ی بهتری نمی‌دهد اما حجم داده بسیار کم بوده و مدرک محکم‌تری نیاز است. هم‌چنین افزایش وزن توانایی‌های تکنیکال بیشترین تاثیر را در درستی دارد.

- (۲۰۱۹) توسعه‌ی فریمورک برای انتخاب کاندیداهای شغلی [3]

از روش ساده‌ی شمارش تعداد توانایی ها استفاده می‌کند. در مورد پردازش متون، از کتابخانه SpaCy استفاده شده و هربار اگر بازخورد بگیرد که یک تشخیص نادرست بوده، در مرتبه‌ی بعدی آن را درست می‌کند. نرخ درستی در حدود ۹.۰ تا ۱ بوده (تعداد داده=۱۵۰ بسیار کم بوده و در حلقه‌ی بازخورد[1] افتاده). سپس برای درخواست بهترین رزومه مطابق با یک شغل، تنها کوئری‌هایی با محتوای وجود تعدادی ویژگی خاص نوشته شده و روی زمان پاسخ آن‌ها بحث شده. در مورد اینکه چه قدر پیشنهاد خوبی بوده، صحبتی نشده.

[1] Feedback loop

- PROSPECT (۲۰۱۰) سیستم نظارت متقاضیان برای استخدام‌کنندگان [4]

این مقاله توسط IBM نوشته شده و ارجاعات بسیار زیادی به این مقاله داده شده است. از همان TF-IDF استفاده کرده. در وبسایت مربوطه برای استخدام‌کننده امکان انتخاب فیلترهایی را گذاشته که همه‌ی رزومه‌های دارای آن مقاله می‌بایست نشان داده شوند. رزومه‌های تکراری را در نظر گرفته تا اگر برای یک بار پذیرفته نشدند، دیگر در نظر گرفته نشوند. از ابزار Lucene[1] و POI[2] برای استخراج متن و اطلاعات از فایل‌های PDF یا Word استفاده شده و در ادامه‌ی دقیق در مورد جزئیات این عمل توضیح داده شده است. هم‌چنین از مدل آماری Conditional Random Fields و مدل مارکف برای استخراج متن کمک می‌گیرد.

به عنوان ورودی یک جایگاه شغلی و تعداد زیادی رزومه دریافت می‌کند و به عنوان خروجی به ترتیب شبیه‌ترین رزومه‌ها را می‌دهد. این امکان را دارد که استخدام کننده فیلترها و شرایطی را روی خروجی اعمال کند. از ویژگی‌های متفاوت این سیستم، اول تشخیص رزومه‌ها و افراد تکراری و دوم، نحوه‌ی نمایش نتایج است که در صفحه‌ی بعد می‌بینید. برای کلاس‌بندی رزومه‌ها از SVM استفاده می‌کند. یکی از توانایی‌های مهم این سیستم پیشنهاد استعداد در هنگامی است که کارفرما به دنبال استعدادهای خاصی است. برای آزمایش نهایی ۸ جایگاه شغلی در نظر گرفته شده و برای هر کدام ۲۰ رزومه از میان ۲۰۰۰ رزومه پیشنهاد داده شده است.

	K	okapi BM25	KL	مدل امتیازدهی Lucene	مدل امتیازدهی +Lucene در نظر گرفتن توانایی ها
دقت k تایی	۵	۰.۱۸	۰.۴۳	۰.۶۰	۰.۷۰
	۱۰	۰.۱۹	۰.۳۴	۰.۵۹	۰.۶۳
	۲۰	۰.۲۱	۰.۳۶	۰.۴۹	۰.۶۱
NDCG	۵	۰.۱۵	۰.۴۴	۰.۵۸	۰.۷۳
	۱۰	۰.۱۷	۰.۳۷	۰.۵۷	۰.۶۶
	۲۰	۰.۱۹	۰.۳۸	۰.۵۱	۰.۶۴

جدول ۱- مقایسه‌ی روش‌های مختلف برای رتبه‌بندی کاندیداهای استخدام [14]

[1] https://lucene.apache.org/
[2] https://poi.apache.org/

	K	مدل امتیازدهی Lucene+ در نظر گرفتن توانایی ها	مدل امتیازدهی Lucene+ توانایی ها+سوابق تحصیلی
دقت k تایی	۵	۰.۶۰	۰.۸۰
	۱۰	۰.۵۰	۰.۶۵
	۲۰	۰.۴۸	۰.۶۵
NDCG	۵	۰.۶۸	۰.۸۶
	۱۰	۰.۵۹	۰.۷۳
	۲۰	۰.۵۴	۰.۷۰

جدول ۲- ارزیابی رتبه‌بندی کاندیداها با در نظر گرفتن سوابق کاری و بدون آن [14]

نتایج با استفاده از NDCG و Precision @k در جدول‌های ۱ و ۲ قابل مشاهده است. همان‌طور که مشاهده می‌کنید با در نظر گرفتن تعداد سال سوابق تحصیلی نتایج بهتری به دست آمده است. فرمول مورد استفاده برای نتایج از اشخاص واقعی استفاده می‌کند.

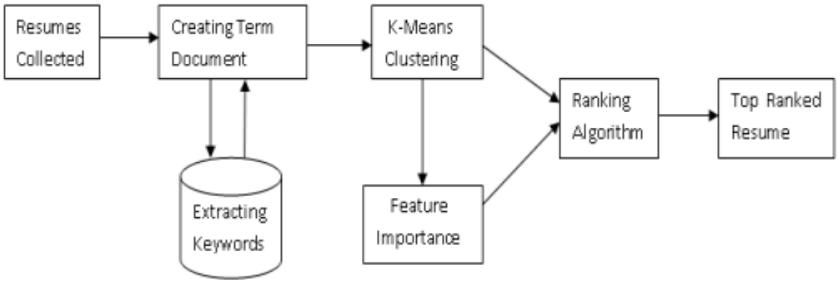
- (۲۰۱۸) سیستم پیشنهادگر شغل بر اساس محتوا[1] با استفاده از Dense Representation [5]

تحقیقاتی روی یافتن شباهت بین متون و ابزار تعبیه‌کردن جملات و یافتن بهترین الگوریتم تعبیه‌سازی انجام شده. بهترین ابزار با استفاده از آزمایش‌ها Doc2vec با ابعاد ۵۰۰ معرفی شده. برای سیستم جست و جوگر از Elastic Search استفاده شده. هم‌چنین فاصله برحسب کیلومتر بین شرکت مورد نظر و متقاضی در نظر گرفته شده.

- (۲۰۱۷) رتبه‌بندی رزومه‌ها بر اساس خوشه‌بندی برای بهبود سیستم استخدام به وسیله‌ی متن‌کاوی و یادگیری ماشین [6]

[1] Document-based recommender system

رتبه‌بندی بر اساس خوشه‌بندی برای هدف استخراج اطلاعات مهم از رزومه‌ها و رتبه‌بندی آن‌ها بر اساس مشابهت



شکل ۴.۳- معماری سیستم در مقاله‌ی [6]

به یکدیگر است. از میان همه‌ی رزومه‌ها آن‌هایی را که تعداد مشخصی از کلمات مهم را داشتند، جمع آوری شده (همه در حوزه‌ی IT هستند).

از شمارش عادی کلمات استفاده شده و ماتریسی با ابعداد رزومه تعداد کلمات مهم ساخته شده (ابعاد بسیار بزرگ). همان‌طور که در شکل ۴.۳ مشاهده می‌کنید از الگوریتم KMeans برای دسته‌بندی و فاصله‌ی اقلیدسی استفاده شده. در کل رزومه‌ها در ۱۰ دسته جدا شده‌اند و ویژگی‌های این ۱۰ دسته استخراج شده اند.

- (۲۰۱۳) سیستم هوشمند EXPERT برای انتخاب استعداد[7]

به جای رزومه از CV استفاده شده (میانگین و جزئیات سوابق را داراست). از یک فرمول برای پیدا کردن شباهت بین رزومه و توضیحات استفاده کرده و نه تکنیک‌های یادگیری ماشین. در نگاهی به آینده می‌خواهد با استفاده از لیست دوستان شخص در شبکه‌های اجتماعی تصمیم‌گیری کند.

- (۲۰۱۸) درست کردن لیست کوتاه از بین رزومه‌های ارسالی [8]

این مقاله هم تنها از محاسبه‌ی یک فرمول برای پیدا کردن میزان شباهت استفاده می‌کند. اما برای بهبود پیشنهاد، از نتیجه‌ی انتخاب شدن یک رزومه در گذشته هم به عنوان وزن استفاده می‌کند.

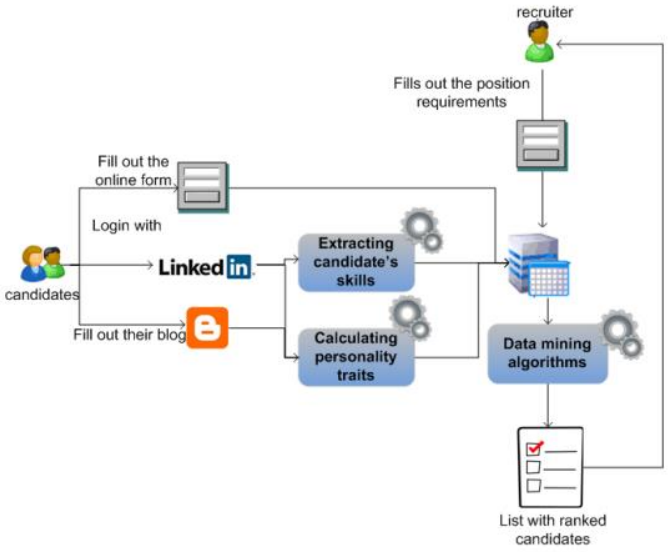
- Applicant Tracking System (ATS) (۲۰۱۹)

پلتفرم‌هایی با این عنوان وجود دارند که روند استخدام، از شروع انتشار جایگاه شغلی تا تنظیم کردن ساعت مصاحبه را برای شما ساده می‌کند. روی کارمندان را به طور خودکار انجام می‌دهند؛ به عنوان مثال در حوزه ی استخدام، ابتدا رزومه‌های ورودی را به فرمت خاصی در می‌آورند و اگر با معیارهای توضیحات جایگاه شغلی تا

حدی سازگاری داشت، آن را در صف نشان دادن به استخدام‌کننده قرار می‌دهند. در این وبسایت[1] تعدادی از آن‌ها معرفی شده اما در مورد پیاده‌سازی آن‌ها اطلاعی نداریم؛ هم‌چنین ممکن است که تنها از یک جست و جوی ساده برای وجود یک استعداد خاص در رزومه استفاده کنند. به عنوان مثال یکی از معروف‌ترین برنامه‌ها Taleo[2] و GreenHouse[3] هستند که توسط بسیاری از شرکت‌های بزرگ استفاده می‌شود.

- (۲۰۱۲) کاربرد یادگیری ماشین در سیستم استخدام آنلاین [9]

سیستمی برای شرکت‌ها طراحی شده تا کارمندان را قبل از استخدام به خوبی شناسایی کنند. بر ویژگی‌های سابقه‌ی تحصیلی، تجربه کاری، علاقمندی‌ها و میانگین کار در هر شغل مد نظر گرفته شده. سیستم دارای بخش ثبت نام است در سیستم که می‌تواند توسط لینکدین انجام شود. نوشته‌های افراد را در این شبکه اجماعی تحلیل



شکل ۵.۳ معماری سیستم در مقاله‌ی [9]

می‌کند. محاسبه‌ی نمره‌ی افراد با توجه به شباهت به جایگاه شغلی که درخواست می‌دهند انجام می‌شود.پس از انجام این عملیات از بین افرادی که برای هر جایگاه اقدام کرده‌اند، برترین‌ها برای مصاحبه انتخاب می‌شوند.

از پارامتر LIWC = linguistic inquiry word count برای تحلیل صفحه‌ی لینکدین افراد استفاده شده.

---

[1] https://www.capterra.com/applicant-tracking-software/

[2] https://www.oracle.com/human-capital-management/taleo

[3] https://www.greenhouse.io

همان‌طور که در شکل ۵.۳ هم نمایش داده شده برای ثبت نام صفحه‌ی لینکدین استفاده می‌شود یا کاربر خود اطلاعاتش را وارد می‌کند. سابقه شغلی در کارمندان قدیمی‌تر و برای کارمندان تازه‌کار، تحصیلات بیشترین تاثیر را دارد. از بین روش‌های درخت تصمیم و SVM، درخت تصمیم بهترین نتیجه را برای پیش بینی داده است.

۲.۲.۳ مرتبط ساختن چند رزومه و چند درخواست

- در وبسایت کگل دو دیتاست معروف[1] در این زمینه وجود دارد که به عنوان مسابقه در این سایت گذاشته شده‌اند. در این پیوندها می‌توانیم تعدادی از نوتبوک‌های پایتون را ببینیم که به تحلیل این سری داده‌ها و پیش‌بینی برای داده‌های تست پرداخته اند.

- CASPER (۲۰۰۰) [10]

جزو اولین مقالات در این حوزه و از معروف‌ترین‌های آن‌هاست و تقریبا همه‌ی مقالات جدیدتر به این مقاله ارجاع داده‌اند. بر جست و جوی بر حسب لغات در حیطه‌ی یافتن شغل ایراد وارد کرده و قصد طراحی سیستمی هوشمندانه‌تر دارد.

دو سیستم به نام ACF و PCR معرفی می‌کند؛ اولی برای دنبال کردن رفتار کاربر و تشکیل یک پروفایل از فعالیت‌های او و دومی که بخش اول آن شباهت‌ها را جست و جو می‌کند و بخش دوم که بر اساس شباهت‌ها به کاربران شغل پیشنهاد می‌دهد. این بخش از پیدا کردن شباهت بین کاربران موجود در سایت و نه بر اساس محتوا استفاده می‌کند. و برای پیدا کردن مشابه‌ترین‌ها، از الگوریتم KNN بهره می‌برد. در زمان خود نتایج بسیار خوبی را یافته است.

- مقاله‌ی A Language Model based Job Recommender (۲۰۱۵) [11]

هدف از این مقاله نوشتن الگوریتمی برای مرتبط ساختن تعدادی زیادی شغل به تعدادی رزومه است. یکی از مشکلات بیان شده در این مقاله است این که: فرض کنید شخصی که مدتی در یک حوزه کار کرده و رزومه‌ی او بر این اساس نوشته شده، بخواهد وارد کار جدیدی شود که متفاوت با قبلی است اما توانایی‌اش را دارد. حال اگر

[1] https://www.kaggle.com/kandij/job-recommendation-datasets و https://www.kaggle.com/c/job-recommendation

سیستم تنها بر اساس لغات موجود در رزومه‌ی او تصمیم بگیرد که شغلی برای او مناسب است، مشکل پیش می‌آید.

از TF-IDF استفاده کرده. مدل‌های ریاضی متفاوتی را برای محاسبه‌ی میزان شباهت و هماهنگی رزومه و یک شغل امتحان کرده. به دلیل حجم بالا از ابزار MapReduce در بحث کلان‌داده استفاده کرده. سپس در مورد معماری برنامه و جزئیات پیاده‌سازی توضیح داده شده. پس از همه‌ی تلاش‌ها برای ساخت بهترین سیستم پیشنهادگر، گفته شده می‌توان با استفاده از بازخورد کاربران، سیستم را روز به روز بهبود بخشید.

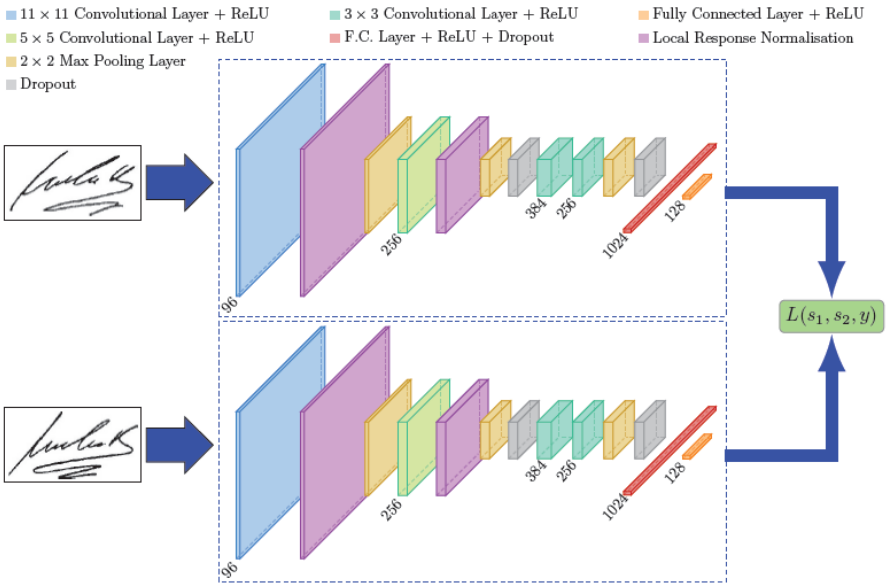
- (۲۰۱۷) پیدا کردن مدل جمعی برای اتصال توضیحات شغل‌ها به رزومه‌های افراد[12]

نقطه قوت: حجم داده در حدود چند میلیون است و از دیتابیس‌های ABG و QABA استفاده شده. ظرایف ریاضی و فرمول مناسب برای میزان شباهت به خوبی توضیح داده شده است.

از روش پیشنهاد بر اساس محتوا[1] استفاده کرده که همان Cosine Similarity است هم‌چنین این روش را با پیشنهاد جمعی[2]مقایسه کرده و در ادامه در مورد تعریف فرمول میزان شباهت بحث شده.

- (۲۰۱۸) استفاده از شبکه‌ی عصبی Siamese برای پیدا کردن ارتباط بین رزومه‌ها و شغل‌ها [13]

[1] Content-based recommendation
[2] Collaborative filtering recommendation



شکل ۶.۳ شبکه‌ی عصبی Siamese [13]

در مقدمه‌ی این مقاله نوشته شده این شبکه‌ی عصبی باعث می شود موارد مشابه به هم نزدیک‌تر و مواردی که ارتباط کمتری دارند، از هم دورتر شوند. در شکل ۶.۳ می‌توانید تصویر معماری این شبکه را مشاهده کنید. این شبکه‌ی عصبی برای مواردی طراحی شده که حجم داده به اندازه‌ی کافی بزرگ نیست. همانطور که مشاهده می‌کنید این شبکه دقیقا از دو بازوی کاملا مشابه تشکیل شده و در بخش اتصال، می‌تواند میزان شباهت دو داده‌ی ورودی را پیدا کند. هم‌چنین این توانایی را دارد که برای تست ورودی با کلاس جدید، نیازی نیست تا دوباره همه‌ی داده‌ی آموزش از اول به آن داده شود.

ادعا شده این مقاله از مقاله‌های مشابه دیگر موجود بسیار بهتر عمل کرده است.

- (۲۰۱۹) سیستم پیشنهادگر شغل در حوزه‌ی فناوری اطلاعات [14]

در مورد چالش‌های پیدا کردن شغل در وبسایت‌های مخصوص این کار بحث می‌کند. به عنوان مثال بیان شده ممکن است برای یک جست و جو، با توجه به حجم زیاد داده تعداد زیادی شغل نشان داده شود و کاربر را گیج کند. هم‌چنین گفته شده بهتر است اگر یک کاربر جست و جوهای متفاوتی انجام می‌دهد، در نتایج جست و جوهای جدید، داده‌های گذشته هم دخیل باشند. ابتدا از TF-IDF برای پارامتریزه کردن متون استفاده کرده اما سپس به دلیل بازدهی پایین به سمت LSA رفته و از روش‌های ماتریسی برای کاهش ابعاد داده استفاده کرده

است. سپس از فرمول‌های مشابهت مانند کسینوس استفاده شده تا ارتباط و نزدیکی بین خود کاربران با هم، و شغل‌ها با هم مشخص شود.

روشی به نام الگوریتم 3A معرفی شده. این الگوریتم از گرافی جهت‌دار و وزن‌دار برای ایجاد ارتباط و پیدا کردن مشابهت‌های بین شغل‌ها،کمپانی‌ها و افراد استفاده می‌کند. حال در مقایسه‌ی روش‌های مختلف برای فیلترینگ، ترکیب نوع بر اساس محتوا و ارتباط با دیگر کاربران بهترین نتیجه را داده است.

۳.۲.۲ یک روزمه و چند شغل

- (۲۰۲۰) پیشنهاد شغل بر اساس رفتار متقاضی و دسته‌بندی پروفایل‌های کاری[15]

توضیحات دقیقی درباره‌ی جزئیات داده نشده اما در نظر داشته از KMeans استفاده کند. پیشنهاد شده به جای استفاده از تمرکز بر روی پیشنهاد دهنده‌ی کار و متقاضی، از روش فیلترینگ بر مبنای مشارکت استفاده شود. یعنی بین اشخاص مشابه پیدا شوند و انتخاب‌های آن‌ها به یکدیگر پیشنهاد داده شود.

- (۲۰۲۰) پیشنهاددهنده‌ی شغل با استفاده از یادگیری ماشین و پردازش زبان‌های طبیعی[16]

به مسأله‌ی شروع سرد[1] اشاره شده؛ به این معنا که هنگامی که کاربر وارد پلتفرم پیشنهاد شغل وارد می‌شود، چون هنوز هیچ پیشینه‌ای ندارد، سیستم نمی‌تواند به او پیشنهاد مناسبی بدهد؛ پس تنها بر متن خود رزومه اکتفا کرده.

از وبسایت Stackoverflow.com برای تأمین داده استفاده شده، که در هیچ یک از مقالات دیگر نبوده. در این مقاله هم میان NLTK و SpaCy،SpaCy به دلیل کارایی و سرعت و دقت بیشتر انتخاب شده. از میان روش‌های encoding موجود یعنی CBOW و Skip-Gram مورد دوم انتخاب شده.

از فرمول شباهت کسینوس برای پیدا کردن شباهت متون استفاده کرده و نرخ درستی در حدود ۶.۰ بوده است.

- (۲۰۱۱) پیشنهاد شغل به افراد شاغل که قصد تغییر شغل دارند، بر اساس تاریخچه قبلی[17]

[1] Cold Start

این مقاله سعی در پیش‌بینی آینده‌ی شغلی افراد دارد و این نتایج را به آن‌ها به عنوان پیشنهاد نشان می‌دهد. این مقاله ادعا می‌کند این کار به راحتی انجام شدنی است و می‌گوید بهترین داده در تاریخچه‌ی یک شخص به عنوان مثال تحصیل یا سابقه‌ی کاری گذشته، تنها همان شغل کنونی اوست. مدل سابقه‌ی شغلی اشخاص را بر حسب یک گراف تعریف کرده. از بین الگوریتم‌های معروف موجود، Naive Bayes بهترین نتیجه را داده است. سه سناریو برای انجام آزمایش در نظر گرفته شده و در حالتی که تنها پیش‌بینی در داده‌های ۱۰۰ کمپانی برتر انجام شده، بهترین نتیجه را داده است (۷۸٪).

- (۲۰۱۳) سیستم پیشنهاددهنده‌ی شغل برای متقاضیان و وبسایت‌های استخدام (iHR (Hybrid [18]

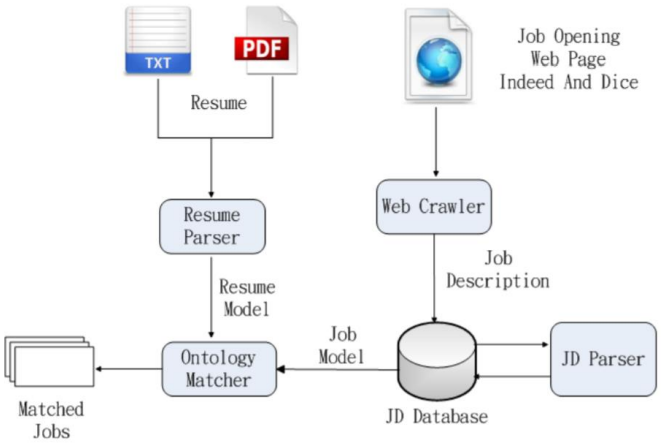
با استفاده از ترکیب اطلاعات خود اشخاص و شباهت میان رفتار افراد، و الگوریتم 3A یک سیستم طراحی می‌کند. قبل از آن به مقایسه‌ی چند سایت برتر در این زمینه یعنی PROSPECT-eRecruiter-CASPER-Proactive می‌پردازد.

- (۲۰۱۷) سیستم پیشنهادگر شغل بر اساس موضوعات[19]

این مقاله استعدادهای کاربران را تحلیل می‌کند و هم متون جایگاه‌های شغلی را و سپس بین این دو شباهت پیدا می‌کند. برای پیش‌پردازش از روش LDA[1] استفاده می‌کند. این روش این مزیت را دارد که در فضای حافظه ثابت اجرا می‌شود و با افزایش حجم داده، محدودیتی ایجاد نمی‌کند. از انواع bag of word است و ترتیب برای آن اهمیتی ندارد. از انواع یادگیری ماشین بدون برچسب است و می‌تواند از چند متن، موضوعات مختلف آن‌ها را خارج کند. برای شباهت از شباهت کسینوس و flexible string matching استفاده شده است. سپس ماتریس مشابهت برای سطر کاربران و ستون شغل‌ها تشکیل می‌شود. در انتها برای پیشنهاد شغل چون بین کاربران و شغل‌ها اشتراک وجود داشت، از TF-IDF بهره برده شده . درستی محاسبه شده دراین مقاله ۹۵٪ است.

- (۲۰۱۶) پروژه‌ی ResuMatcher [20]

[1] از میان جایگاه‌های شغلی مختلف به یک لیست کوتاه میرسد تا در زمان صرفه جویی کند.



شکل ۷.۳- معماری سیستم در مقاله‌ی [20]

مسأله در این مقاله پیدا کردن بهترین شغل برای یک رزومه‌ی یک شخص است و می‌خواهیم مشابه‌ترین متن یک فرصت شغلی را برای یک رزومه پیدا کنیم. همان‌طور که در شکل ۷.۳ مشخص شده، ورودی هم از طریق فایل است و هم از طریق وبسایت کارجویی. خروجی تعدادی رزومه به ترتیب شباهت هستند.برای گرفتن رزومه‌ها از کاربر اینکه از او بخواهیم در فیلدهای مخصوصی اطلاعات خود را وارد کند، کار زمان‌بری است پس استخراج داده به صورت مستقیم از رزومه افراد مد نظر قرار داده شده. داده‌ها از وبسایت Indeed.com و در حوزه‌ی IT درخواست داده شده اند و رزومه افراد در قالب HTML است نه pdf. برای پردازش زبان طبیعی از NLTK بهره برده شده و برای مشابهت برچسب‌های متن از FST استفاده شده.

K	دقت k تایی		DCG	
	Indeed	ResuMatcher	Indeed	ResuMatcher
۵	۰.۸۴	۰.۸۷	۲۳.۸۷	۳۲.۹۷
۱۰	۰.۷۲	۰.۸۶	۳۷.۰۲	۴۵.۵۷
۲۰	۰.۶۵	۰.۷۶	۵۸.۷۰	۶۶.۷۰

جدول ۳- مقایسه‌ی نتایج برای میانگین شغل‌های جاوا و پایتون در سایت indeed و Resumatcher [20]

در جدول ۳ می‌توانیم نتیجه را برای میانگین امتیاز رزومه‌های پیشنهادشده برای شغل های مربوط به حوزه‌ی جاوا و پایتون مشاهده کنیم. برای هر دو پارامتر مشخص شده، این سیستم بهتر از وبسایت عمومی indeed عمل کرده است.

۴.۲.۳ یک رزومه و چند شغل

- (۲۰۱۹) پیشنهاد رزومه با استفاده از شبکه ی عصبیBERT

این مقاله یک طرفه خواندن متون توسط مدل‌های قبلی را جبران می‌کند و دو طرفه متون را میخواند. برنامه استفاده از آن سبک پیاده سازی شده. بر خلاف W2V GLOVE کلمات را با توجه به موضوع جمله‌ی آن ها وکتوریزره می‌کند.

- (۲۰۱۵) پیشنهاد شغل در شبکه‌های اجتماعی مانند فیسبوک بر اساس پروفایل و رفتار افراد

درباره‌ی انواع سیستم‌های پیشنهادگر بر اساس رفتار کاربر و شباهت او به کاربران دیگر توضیح داده شده. راهکارهایی برای استخراج شخصیت کاربر از روی رفتار و محتوای تولیدی توسط او داده شده. برای قسمت یادگیری ماشین از الگوریتم SVM و شبکه‌های عصبی عمیق استفاده شده است. مقاله بسیار دقیق اما متفاوت با این پروژه است.

- (۲۰۲۰) استخراج توانایی‌ها از یک توضیحات فرصت شغلی [22]

ادعا شده ۶۵٪ توضیحات شغل‌ها در توضیح توانایی‌های مورد نیاز خود به خوبی موفق نبوده‌اند و می‌خواسته این مشکل را حل کند. هم‌چنین ادعا شده در ۴۰٪ رزومه‌های دیتاست مربوطه، ۲۰٪ توانایی‌ها در توضیحات نوشته شده و نه در لیست نیازمندی‌های توانایی. این مقاله برای داشتن لیستی کامل از توانایی‌ها بر حسب توضیحات ابتدایی شغل، از الگوریتم‌های کلاس‌بندی‌های چندگانه (با چند برچسب) و شبکه‌ی عصبی BERT استفاده کرده است.

- (۲۰۱۸) تحقیقی بر سیستم‌های پیشنهادگر شغلی [23]

گفته شده تحقیقات زیادی در زمینه‌ی سیستم پیشنهادگر شغل در سال‌های اخیر انجام شده. اما میان آن‌هایی که رزومه‌ی شخص را به شغل مربوط می‌سازند و آن‌هایی که رفتار کاربر را در شبکه‌های اجتماعی هم در نظر می‌گیرند، مورد اول بهترین نتیجه را داشته (اطلاعات کمتر، نتیجه بهتر).

در یک آزمایش بر اساس پروفایل تعدادی از کاربران، با استفاده از ۴ روش متفاوت یک پوستر فرصت شغلی ساخته‌اند و در مقابل کاربر قرار داده تا او یک فرصت شغلی را انتخاب کند. در این آزمایش روش Word2vec-Skip-Gram بهترین نتیجه را با ۵۹.۰ درصد درستی داده است.

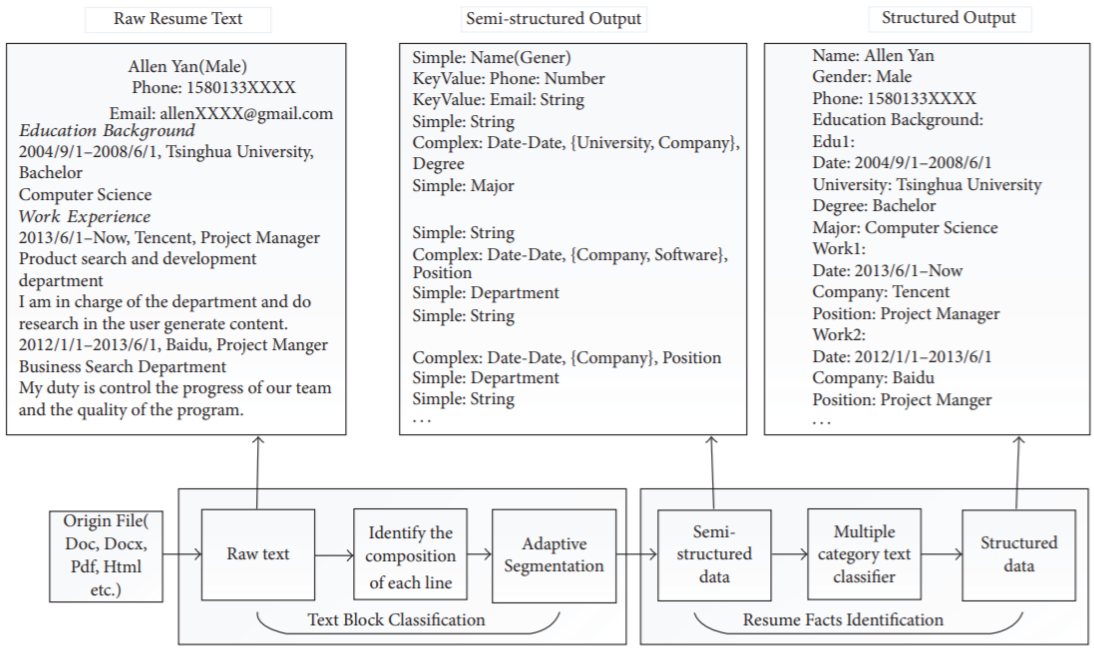
- (۲۰۱۹) پیدا کردن ارتباط بین رزومه‌ها و شغل‌ها بر اساس مکان در نقشه [24]

با استفاده از کتابخانه‌ی Beautiful Soup در پایتون شغل‌ها را از وبسایت‌های معروف استخراج کرده و با استفاده از اطلاعات متون رزومه و مکان جغرافیایی، بردارهایی تهیه شده که شباهت آن‌ها با کمک Cosine Similarity به دست آمده. برای پیشنهاد به کاربر از میان مشابه‌ترین متون به رزومه‌ی فرد آن شغل‌هایی را که در یک محدوده‌ی جغرافیایی وجود دارند بر روی نقشه به او پیشنهاد می‌دهد.

- (۲۰۱۲) ارزیابی سیستم‌های پیشنهاد شغل الکترونیکی [25]

در این مقاله ابتدا بیان شده روش Boolean Search روش مناسبی نیست زیرا در این حالت اهمیت توانایی‌ها یا ارتباط بین آن‌ها در نظر گرفته نمی‌شود و بهتر است از روش‌های پیچیده‌تر استفاده کرد. ۶ نوع سیستم‌های استخدام توضیح داده شده که می‌توان ساختار همه‌ی آن‌ها را با اندکی تفاوت مشابه دانست. سپس به مقایسه و بررسی انواع چالش‌های سیستم‌های پیشنهادگر پرداخته شده است.

- ✓ (۲۰۱۸) الگوریتم دو مرحله‌ای برای استخراج اطلاعات از رزومه [26]



شکل ۸.۳-معماری سیتم مقاله‌ی [26]

هدف به دست آوردن درست متون از فایل‌های پی‌دی‌اف یا ورد، و سپس تبدیل درست متون به مقادیر برچسب‌گذاری شده است. مطابق شکل ۸.۳ ، ابتدا مقادیر به صورت خام از فایل‌های ارسالی بازیابی شده، سپس به صورت اطلاعات «کلید-مقدار» مشابه شماره تلفن، ایمیل، ... یا «ساده» یعنی اطلاعات تیتروار یا «پیچیده» یعنی متن بلند تبدیل می‌شود و بعد از آن برچسب‌گذاری می‌شود(Text Classification با استفاده از Naive Bayes). سپس با استفاده از مشابهت کسینوس مقادیر به دست آمده از TF-IDF شبیه بودن بلاک‌ها بررسی شده و با استفاده از KMeans برچسب‌گذاری می‌شوند.

فرمول‌های دقت و پوشش برای محاسبه‌ی دقت برنامه استفاده شده است و نشان می‌دهند تا چه حد اطلاعات درستی از رزومه استخراج شده است.

نتیجه‌ی درستی در برخی از موارد مانند تجربه کاری یا اطلاعات شخصی در مقایسه با [4] و CHM (یک ابزار چینی استخراج اطلاعت از متن مربوط به سال ۲۰۰۵) بهتر عمل کرده و در کل مدت زمان کمتری برای اجرا مصرف کرده. بیان می‌کند اگر همه‌ی دادگان در در رابطه با یک نوع شغل بودند، نتیجه‌ی بهتری می‌گرفته (مانند [4])

✓ (۲۰۱۹)استخراج اطلاعات از رزومه با استفاده از الگوریتم جدید بازه‌بندی بلاک‌های متن[27]

Personal	Education	Work	Project	Skill	Publication
Name; Address; Phone; Email;	University; Graduate school; Graduation Date; Major; Degree;	Company; Job title; Work Period; Job Description;	Title; Project Period; Project Description;	Language; Computer ;	Reference;

شکل ۹.۳- قالب اطلاعات استخراج شده توسط مقاله[27]

با استفاده از روش‌های شبکه‌های عصبی عمیق بلوک‌های متن را از فایل‌ها استخراج می‌کند و اطلاعات به فرم شکل ۹.۳ در جدول ذخیره می‌شوند. سپس اطلاعات مهم را از متون استخراج کرده و به وکتوریزه کردن می‌پردازد.

با توجه به جایگاه متون در فایل و شبکه‌های عصبی این استخراج انجام شده است. برای پردازش داده Word2Vec BERT GloVe استفاده شده است.

برای کلاس‌بندی متون از Text-CNN RCNN Adversarial LSTM استفاده شده است.

۳.۳ جدول مقایسه‌ی کارهای پیشین

دسته	ابزار و الگوریتم‌های مشابهت و پیشنهاددهنده	روش‌های تعبیه‌سازی متن	ابزار پیش‌پردازش	نقاط منفی	نقاط مثبت	مرجع
۱	TF-IDF	TF-IDF	NLTK Pandas Numpy	در حد پروژه‌ی آکادمیک و غیر قابل استفاده در واقعیت	استفاده از ابزارهای جدید و به روز به همراه تحلیل داده	[1]
	KMeans KNN	ذکر نشده	ذکر نشده	حجم داده‌ی کم	خلاصه‌سازی توانایی‌های کاربران	[2]
	ذکر نشده	ذکر نشده	SpaCy	استفاده از روش ساده‌ی شمارش توانایی‌ها	ذخیره حافظه برای پیشنهادهای بعدی	[3]

		Feedback loop				
[4]	PDFوWord	ذکر نشده	Lucene POI	TF-IDF	TF-IDF مدل مارکف SVM برای کلاسبندی	
[6]	استخراج حداکثری اطلاعات	همه در حوزه IT شمارش تعداد کلمات مشابه مقیاس ناپذیری	ذکر نشده	ذکر نشده	KMeans Cosine Similarity	
[7]	استفاده از cv به جای رزومه	استفاده نکردن از یادگیری ماشین	ذکر نشده	ذکر نشده	فرمول ابتکاری	
[8]	استفاده از داده‌ی برچسب دار به عنوان رزومه‌ی انتخاب شده یا نشده	جزئیات آورده نشده	ذکر نشده	ذکر نشده	فرمول ابتکاری	
[9]	در نظرگرفتن تمامی ویژگی‌های متقاضیان شغل استفاده از رفتار کاربر در لینکدین یا ثبت نام مستقیم	جزئیات آورده نشده	ذکر نشده	ذکر نشده	SVM درخت تصمیم	
[10]	جزو اولین مقالات تحلیل رفتار کاربر و تشکیل پروفایل برای او	مقیاس ناپذیر	ذکر نشده	ذکر نشده	KNN	۲
[11]	در نظر گرفتن عوض شدن سلیقه‌ی کاربر استفاده از MapReduced برای حجم بالای داده	جزئیات آورده نشده	ذکر نشده	TF-IDF	ذکر نشده	

	سیستم پیشنهاددهنده‌ی مشارکت محور شباهت کسینوس	ذکر نشده	ذکر نشده	جزئیات آورده نشده	استفاده از دیتابیس‌های بزرگ برای آموزش مدل آوردن روش‌های ابتکاری ریاضی	[12]
	Siamese	Word2vec	ذکر نشده	جزئیات آورده نشده	روش ابتکاری برای پیشنهاد	[13]
	شباهت کسینوس	TF-IDF	ذکر نشده	تنها در حوزه‌ی IT است	استفاده از الگوریتم 3A در نظرگرفتن سابقه‌ی جست و جوهای یک کاربر	[14]
۳	روش پیشنهاددهنده‌ی مشارکت محور KMeans	ذکر نشده	ذکر نشده	جزئیات آورده نشده	پیدا کردن مشابهت بین افراد مختلف	[15]
	شباهت کسینوس	CBOW, SKipGram	مقایسه NLTK و SPaCy	جزئیات آورده نشده	توجه به شروع سرد استفاده از منبع داده‌ی stackoverflow	[16]
	Naive Bayes	ذکر نشده	ذکر نشده	جزئیات آورده نشده	پیش‌بینی شغل آینده افراد	[17]
	روش ابتکاری بر اساس محتوا و الگوریتم 3A	ذکر نشده	ذکر نشده	جزئیات آورده نشده	استفاده از رفتار و پروفایل کاربر	[18]
	Cosine similarity Flexible String Matching	Bag of Word TF-IDF	ذکر نشده	جزئیات آورده نشده	استفاده از حافظه‌ی کم حین اجرا	[19]
	روش ابتکاری	روش ابتکاری	NLTK	تنها برای فایل html تنها در حوزه‌ی IT	ارائه روش‌های ابتکاری ارائه در قالب وبسایت	[20]

جدول ۴- مقایسه‌ی مقاله‌های پیشین

۴.۳ نتیجه گیری

با توجه به مقایسه‌ی کارهای پیشین انجام شده، در سیستمی که طراحی می‌کنیم، قصد داریم با دو زبان بتوانیم پیشنهاد بدهیم که در هیچکدام از کارهای گذشته این توانایی دیده نشد. هم‌چنین در این تحقیق قصد داریم

روش‌های ماتریسی و مقیاس‌ناپذیر را نادیده بگیریم و از جدیدترین روش‌های یادگیری ماشین برای طراحی سیستم پیشنهاد دهنده استفاده کنیم.

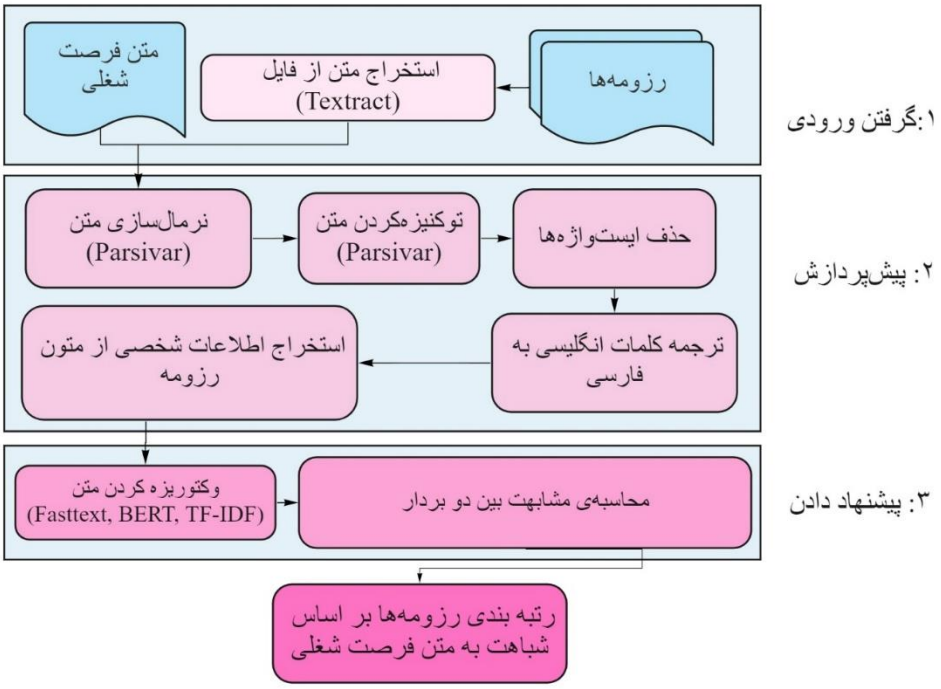
فصل چهارم: رویکرد پیشنهادی

۱.۴ مقدمه

رویکرد پیشنهادی ما برای یک سیستم پیشنهاددهنده‌ی رزومه سعی دارد ابتدا متون را از فایل‌های رزومه با فرمت‌های مختلف استخراج می‌کند. سپس، سعی می‌کنیم بیشترین اطلاعات موجود را از جمله شماره تلفن، ایمیل، لینک‌های مختلف، پروژه‌های دانشگاهی، کار در شرکت و مقطع تحصیلات از متن خام رزومه استخراج و در جدولی ذخیره کنیم. سپس، با استفاده از روش‌های مختلف تعبیه‌سازی متون متن رزومه و فرصت شغلی را به برداری از اعداد تبدیل کرده، و فواصل (یا شباهت) بین هر رزومه و فرصت شغلی را محاسبه می‌کنیم. سپس رزومه‌ها را بر اساس این فواصل (یا شباهت) مرتب کرده، و به استخدام‌کننده نمایش می‌دهیم. در فصل بعدی نتایج حاصل از پیاده‌سازی برای ارزیابی و اعتبارسنجی مورد بررسی قرار می‌گیرد.

در ادامه در بخش ۴-۲ گام‌های رویکرد پیشنهادی توضیح داده شده است. سپس در فصول ۱.۲.۴ تا ۴.۳.۲ هرکدام از بخش‌ها توضیح داده می‌شوند و در انتها در فصل ششم نتیجه‌گیری فصل ارائه می‌شود.

۲.۴ گام‌های رویکرد پیشنهادی



شکل ۱.۴ معماری پیشنهادی سیستم پیشنهاددهنده‌ی رزومه

شکل ۱.۴ گام‌های طی‌شده از خوانش رزومه و متن فرصت شغلی تا رتبه‌بندی نهایی را نشان می‌دهد. این گام‌ها به طور کلی شامل سه بخش ورودی، پیش‌پردازش و پیشنهاد دادن می‌شوند. در بخش ورودی اطلاعات از روی فایل‌ها خوانده می‌شود، در پیش‌پردازش متن را در اصطلاح تمیز می‌کنیم و آن را برای تبدیل شدن به بردار آماده می‌کنیم. در بخش پیشنهاددادن، هر روزمه تبدیل به برداری با طول ثابت شده و با روش‌های مختلف شباهت بین دو بردار را محاسبه می‌کنیم. در نهایت، رزومه‌های برای هر متن فرصت شغلی به ترتیب مرتب شده و به کاربر نشان داده می‌شوند. در ادامه، به بررسی دقیق هرکدام از این بخش‌ها می‌پردازیم.

۱.۲.۴ ورودی

نوع فایل‌هایی که به این برنامه ورودی داده می‌شوند، ممکن است PDF یا Word باشند. فایل‌های Word به گونه‌ای ذخیره می‌شوند که می‌توان متن را بدون اشتباه به راحتی از فایل خواند و این کار برای زبان فارسی و انگلیسی تفاوتی ندارد. اما در مورد فایل‌های PDF، چون متن به صورت خاصی ذخیره می‌شود باید از روش‌های دیگری استفاده کرد. به این منظور تعداد زیادی از ابزارهای استخراج متن در زبان پایتون آزموده شدند. در جدول ۵ لیستی از آن‌ها و دلیل انتخاب شدن یا نشدن هرکدام را آورده‌ایم.

نام ابزار	نقاط مثبت و منفی
Hooshvare pdf2word[1]	ستون متن را متوجه نمی‌شود. اعداد فارسی را نمی‌خواند. ابتدا به فایل ورد و سپس متن تبدیل می‌شود.
PLUMBER[2]	کلمات انگلیسی به درستی استخراج می‌شوند و کلمات فارسی به جز اعداد اگر برعکس شوند، درست نشان داده می‌شوند. مشکل خوانش ستون متن دارد. می تواند جدول‌ها را استخراج کند.
Textract from AWS[3]	قابلیت خوانش فایل‌های متفاوت را دارد. بدون مشکل ستون متنی را جدا می‌کند. زبان فارسی و انگلیسی را به خوبی می‌تواند جدا کند. مستقیم به متن تبدیل می‌کند و نه به فرمت ورد.
PyPDF2[4]	بسیار زمانبر است. ازدست‌رفت اطلاعات دارد. برخی فایل‌ها را به کل نمی‌خواند.

جدول ۵- مقایسه‌ی ابزارهای استخراج متن از PDF

۲.۲.۴ پیش‌پردازش

این مرحله خود دارای پنج بخش است که در زیر هرکدام را توضیح می‌دهیم.

نرمال‌سازی: به این معناست که همه‌ی کاراکترهای متن استاندارد شود. به عنوان مثال در زبان فارسی ممکن است گاهی به جای «ک» فارسی از « ك» که عربی است استفاده شود. در این صورت می‌بایست فرم عربی را به فارسی تبدیل کرد.

توکنیزه‌کردن: به این معنا که به جای آنکه کل متن را به طور یک دست در یک متغیر داشته باشیم، کلمات را بر اساس فاصله یا علائم نگارشی مانند . و ! جدا می‌کنیم.

[1] https://github.com/hooshvare/pdf2word
[2] https://github.com/jsvine/pdfplumber
[3] https://textract.readthedocs.io/en/stable/
[4] https://pythonhosted.org/PyPDF2/

برای بخش‌های نرمال‌سازی و توکنیزه کردن، دو کتابخانه‌ی پایتون در زبان فارسی به نام ۱- هضم[1] و ۲-پارسی‌ور[2] موجود است. در طی مقایسات انجام شده، پکیج پارسی‌ور با دقت بیشتر، و زمان کوتاه‌تری عمل می‌کند. به همین علت از این پکیج استفاده شد.

حذف ایست‌واژه‌ها: ایست‌واژه‌ها کلماتی در متن هستند که با اینکه جمله بدون آن‌ها معنا پیدا نمی‌کند، اما خود به تنهایی معنای خاصی ندارند. مانند کلمات ربط. این کلمات اگر از متن حذف نشوند، پردازشگر دچار خطا می‌شود. به این منظور، ابتدا لیستی از این کلمات تهیه شد، سپس کلمات بیشتری به طور دستی به این کلمات اضافه شد.

ترجمه: در رزومه‌هایی که حاوی کلمات فارسی و انگلیسی با هم هستند، ممکن است به دلیل این اختلاف زبان، نتوانیم پردازش درستی انجام دهیم. یا به زبان دیگر فرض کنید در رزومه‌ای شخصی توانایی «ورد» و شخص دیگری توانایی «Word» را در بین استعدادهای خود آورده. حال با اینکه این دو کلمه درواقع یکی هستند، اما به دلیل اختلاف کاراکترها درستی آن‌ها توسط برنامه تشخیص داده نمی‌شود. به همین دلیل، می‌بایست کلمات همه به یک زبان باشند. ایده‌ی ابتدایی برای ترجمه، استفاده از پکیج‌های آنلاین ترجمه بود که از سرویس ترجمه‌ی گوگل یا وبسایت‌های دیگر استفاده می‌کردند. در جدول ۶ علت انتخاب نشدن پکیج‌های ترجمه‌ی موجود در زبان پایتون ذکر شده است. در این جدول تنها پکیج‌هایی آمده‌اند که زبان فارسی را می‌شناسند. باقی پکیج‌ها به کلی برای زبان فارسی نوشته نشده‌اند.

پکیج ترجمه	علت انتخاب نشدن
Googletrans[3]	سرعت بسیار پایین به علت آنلاین بودن تعداد درخواست‌های محدود در ساعت برخی کلمات را نمی‌شناسد
Translate[4]	سرعت پایین به علت آنلاین بودن محدودیت روزانه برای درخواست
Deep-translate[5]	برای زبان فارسی به خوبی عمل نمی‌کند.

جدول ۶- مقایسه‌ی ابزارهای ترجمه در پایتون

[1] https://www.sobhe.ir/hazm
[2] https://github.com/ICTRC/Parsivar
[3] https://pypi.org/project/googletrans
[4] https://pypi.org/project/translate
[5] https://pypi.org/project/deep-translator

به علت سرعت پایین، محدودیت درخواست و دقیق نبودن پکیج‌های ترجمه، تصمیم گرفته شد مجموعه‌ی کلماتی از انگلیسی که در رزومه‌های فارسی عمدتا دیده می‌شوند، تهیه شود و ترجمه‌ی آن‌ها آورده شود. بنابراین به جای ترجمه‌ی آنلاین، از ترجمه‌ی آفلاین استفاده شد. در این مجموعه دادگان، چون تنها کلمات مورد نیاز آمده‌اند، زمان جست و جو و فضای اشغالی در آن بسیار کم است.

استخراج اطلاعات: برای استخراج الگوهایی از متن از روشی به نام Regex استفاده می‌شود. با استفاده از زبان Regex می‌توان الگوهایی ساخت تا به عنوان مثال همه‌ی شماره‌ی تلفن‌های موجود در متن خارج شوند. با استفاده از این روش، شماره‌های تلفن، ایمیل‌ها، لینک‌های مختلف مانند لینکدین، مقطع دانشگاهی و پروژه‌های دانشگاهی یا در شرکت‌های مختلف در متن یافته و استخراج شدند.

۳.۲.۴. پیشنهاد دادن

این مرحله شامل دو بخش است: ۱- وکتوریزه کردن متن ۲-پیشنهاد دادن و محاسبه‌ی فاصله یا شباهت بین متون

وکتوریزه کردن متن به معنای تبدیل متن به برداری از اعداد است. این کار به چند روش قابل انجام است. روش اول استفاده از مدل‌های شبکه عصبی از پیش آموزش داده شده است. شبکه‌های موجود آموزش داده شده برای زبان فارسی، BERT و Fasttext هستند که هر دو روی دیتابیس ویکیپدیای فارسی آموزش داده شده اند. مدل Fasttext از روش معمول Word2Vec استفاده می‌کند و شبکه‌ی BERT به گونه‌ای آموزش دیده شده که متن را از هر دو طرف می‌خواند. هر دوی این روش‌ها تست شدند. برای تبدیل متن به برداری از اعداد، ابتدا بردار تک تک کلمات را جداگانه به شبکه‌ی عصبی ورودی می‌دهیم و در خروجی، بردار را تحویل می‌گیریم. سپس از همه‌ی بردارها میانگین می‌گیریم.

$$w_{i.j} = tf_{i.j} \times \log\frac{N}{df_i}$$

$tf_{i.j}$: $number\ of\ occurence\ of\ i\ in\ j$

df_i: $number\ of\ documents\ containing\ i$

N: $total\ number\ of\ documents$

روش دوم استفاده از TF-IDF است که با استفاده از فرمول بالا محاسبه می‌شود. در این فرمول i به کلمات متن و j به متون اشاره دارد.

برای بخش دوم، می‌بایست بین متون فاصله یا میزان شباهت محاسبه کنیم. برای این کار می‌توان از دو روش مشابهت کسینوس یا فاصله‌ی اقلیدسی استفاده کرد. هر دوی این روش‌ها در فصل بعد مقایسه شده‌اند.

در نهایت پس از طی این مراحل، رزومه‌ها بر اساس مشابهت یا فاصله نسبت به متن فرصت شغلی مقایسه می‌شوند و به کاربر نشان داده می‌شوند.

فصل پنجم: ارزیابی و اعتبار سنجی

۱.۵ مقدمه

در این فصل رویکرد پیشنهادی ارزیابی و اعتبارسنجی شده است. در این فصل ابتدا مورد مطالعاتی رویکرد پیشنهادی معرفی و تشریح می‌شود. در بخش ۳.۵ نتایج حاصل از پیاده سازی را تشریح می‌کند. در نهایت در بخش ۴.۵ نتیجه‌گیری و جمع‌بندی ارائه شده است.

۲.۵ مورد مطالعاتی

در این پژوهش ابتدا روی ۵۰ عدد از رزومه‌های فارسی مطالعه انجام شد. سپس به دلیل کمبود منابع آزاد رزومه‌ی فارسی، برای ادامه از ۲۰۰ رزومه‌ی انگلیسی استفاده شده. عمده‌ی موضوعات این رزومه‌ها در حوزه‌ی IT است، اما تعدادی با موضوعات حسابداری و گرافیک نیز به آن‌ها اضافه شده است.

۳.۵ نتایج

در ادامه نتایج برای رزومه‌های فارسی و انگلیسی و برای مدل‌های Fasttext و BERT با استفاده از شباهت کسینوس و فاصله‌ی اقلیدسی، و برای روش TF-IDF ارائه شده است. درصد درستی برای این مدل‌ها با استفاده از روش دقت kتایی که در ادبیات موضوع توضیح داده شد، بیان شده است.

k	BERT		Fasttext	
	شباهت کسینوس	فاصله‌ی اقلیدسی	شباهت کسینوس	فاصله‌ی اقلیدسی
۵	۰.۴۰	۰.۵۲	۰.۸۰	۰.۸۰

۱۰	۰.۵۶	۰.۷۰	۰.۷۶	۰.۷۶
۱۵	۰.۶۱	۰.۷۶	۰.۶۶	۰.۷۰
۲۰	۰.۵۱	۰.۵۵	۰.۶۶	۰.۶۶

جدول ۶- ارزیابی رتبه‌بندی رزومه‌های فارسی با استفاده از روش‌های تعبیه‌سازی متن

در جدول ۶ مشاهده می‌کنیم که بهترین نتیایج زبان فارسی مربوط به مقدار k=۵ و ۱۰ و برای مدل Fasttext بوده است. هم‌چنین نتیجه می‌گیریم اختلاف مقادیر برای فرمول‌های شباهت کسینوس و فاصله‌ی اقلیدسی بسیار جزئی است و اختلاف آن چنانی ندارند.

K	BERT		Fasttext	
	شباهت کسینوس	فاصله‌ی اقلیدسی	شباهت کسینوس	فاصله‌ی اقلیدسی
۵	۰.۶۰	۰.۶۳	۰.۶۱	۰.۶۰
۱۰	۰.۶۱	۰.۶۵	۰.۵۸	۰.۶۱
۱۵	۰.۵۷	۰.۵۳	۰.۵۴	۰.۵۵
۲۰	۰.۶۰	۰.۵۴	۰.۵۵	۰.۴۰

جدول ۷- جدول ۶- ارزیابی رتبه‌بندی رزومه‌های انگلیسی با استفاده از روش‌های تعبیه‌سازی متن

در مقایسه‌ی جدول‌های ۶ و ۷ می‌بینیم بر خلاف انتظار نتایج برای رزومه‌های فارسی بهتر از انگلیسی بوده. هم‌چنین به طور کلی نتایج برای مدل Fasttext بالاتر از BERT است. علت این مسأله این است که مدل BERT بیشتر بر ساختار جملات تاکید می‌کند در حالی که Fasttext روی کلمات و معنای آن‌ها تاکید دارد.

k	TF-IDF	
	انگلیسی	فارسی
۵	۰.۶۰	۰.۸۴
۱۰	۰.۶۴	۰.۸۸

۱۵	۰.۴۹	۰.۷۴
۲۰	۰.۴۶	۰.۵۴

جدول ۸- مقایسه‌ی مدل TF-IDF برای زبان فارسی و انگلیسی

در جدول ۸ مشاهده می‌کنیم نتایج مدل TF-IDF در مقایسه با مدل‌های از پیش آموزش داده شده‌ی شبکه‌ی عصبی بهتر عمل کرده‌اند. این مدل با k = ۱۰ دارای بهترین نتیجه در میان آزمایش‌های قبلی است. علت این اتفاق این است که روش TF-IDF دقیقا بر روی درصد تعداد کلمات مشترک تمرکز می‌کند و توجهی به کلمات قبل و بعد یک کلمه ندارد. اما ایراد این مدل این است که نمی‌تواند توانایی‌های مرتبط را پیدا کند. به عنوان مثال اگر کسی Access Database را در میان توانایی‌های خود داشته باشد و کافرما به دنبال توانایی SQL Server باشد، با اینکه این دو توانایی بسیار مشابهند، اما این روش آن را متوجه نمی‌شود.

۴.۵ نتیجه‌گیری و جمع‌بندی

به طور کلی نتیجه می‌گیریم برای زبان فارسی مدل TF-IDF با ۸۸ درصد درستی و برای زبان انگلیسی مدل‌های از پیش آموزش دیده شده با حدود ۶۰ درصد بهترین نتیجه را دارند. نکته‌ی مورد توجه در این آزمایش این است که ممکن است با تغییر مجموعه دادگان این درصد ها نیز تغییر کنند و نمی‌توان از این اعداد نتیجه‌ی قطعی گرفت. اما تا حدودی می‌توان پیش‌بینی کرد با چه روش‌هایی می‌توان نتیجه‌ی بهتری گرفت. نکته‌ی دوم در مورد مدل‌های از پیش آموزش دیده شده این است که این مدل‌ها روی داده‌ی عمومی ویکیپدیا آموزش دیده شده‌اند و احتمال دارد روی داده‌ی تخصصی در مورد شغل‌ها و توانایی‌های مورد نیاز در محل کار به خوبی عمل کنند. از این رو بهتر است با فراهم کردن حجم زیاد داده، مدل‌هایی بر اساس متون رزومه‌های مختلف را آموزش داد.

فصل ششم: نتیجه گیری

۱.۶ مقدمه

سیستم طراحی شده در این پژوهش سعی دارد فرآیند استخدام در منابع انسانی را سرعت ببخشد. در این پژوهش روشی را معرفی کردیم که می‌توان به وسیله‌ی آن بخشی از فرآیند استخدام را خودکارسازی کرد و به دقت و سرعت این پروسه افزود.

در این فصل در بخش ۲.۶ به سوالات تحقیق پاسخ داده می‌شود، در بخش ۳.۶ دستاوردها نام برده می‌شوند در بخش ۴.۶ به محدودیت‌هایی که در طی این پژوهش با آن‌ها مواجه شدیم توضیح می‌دهیم و در انتها به بخش ۵.۶ به نتیجه‌گیری نهایی و کارهای آتی می‌پردازیم.

۲.۶ پاسخ به سوالات تحقیق

۱. چگونه می‌توان از پرونده‌های رزومه با فرمت‌های مختلف، اطلاعات را به صورت درست استخراج و آن‌ها را به صورت ساختاریافته ذخیره کرد؟

بهترین ابزار برای این کار کتابخانه‌ی Textract پایتون است. پس از تبدیل فایل‌های مختلف به متن خام، با استفاده از الگوهای رجکس می‌توان اطلاعات مورد نیاز را از متن‌ها استخراج و آن‌ها را در جدولی با ساختار مشخص ذخیره کرد.

۲. روش‌های پردازش متون رزومه و آماده‌سازی آن‌ها برای مرحله‌ی پیشنهاد دادن کدام‌ها هستند؟

دو روش موجود برای اینکار استفاده از مدل‌های مبتنی بر شبکه‌های عصبی و به دست آوردن فاصله یا مشابهت کسینوس است، یا یا استفاده از مدل TF-IDF است که نسبت کلمات مشابه در دو متن را محاسبه می‌کند. برای روش اول یا می‌توان از مدل‌های از پیش آموزش داده شده استفاده کرد، یا مدل را با استفاده از داده‌ی دلخواه آموزش داد.

۳. روش‌های متفاوت سیستم‌های پیشنهاد دهنده کدام‌ها هستند و چگونه می‌توان برای رتبه‌بندی رزومه‌ها بر اساس متن آگهی شغلی از آن‌ها استفاده کرد؟

انواع سیستم‌های پیشنهاد دهنده عبارتند از: ۱- بر اساس محتوا، ۲- بر اساس کاربران مشابه و ۳- ترکیب ۱ و ۲. در این پژوهش از روش ۱ استفاده شده است.

۴. و در نهایت چگونه می‌توان بخشی از فرآیند استخدام را به کمک پاسخ سوالات بالا، خودکارسازی کرد؟

با ترکیب مراحل معرفی شده می‌توان متون را جمع آوری، پردازش و تبدیل به مقادیر عددی کرد، سپس مشابهت متون رزومه را با متن فرصت شغلی محاسبه کرد. تفاوت تنها در جزئیات این اعمال است.

۳.۶ دستاوردهای تحقیق

در این فصل به طور جزئی دستاوردهای تحقیق را بیان می‌کنیم.

- استخراج متن فارسی و انگلیسی به طور منظم به همراه ترتیب درست اعداد از فایل PDF.

- استخراج اطلاعات مهم از متن رزومه‌ها و ذخیره‌ی آن‌ها به طور ساختاریافته.
- تبدیل متون به بردارهایی معنادار به طوری که بتوان مشابهت یا فاصله‌ی متون را به دست آورد.
- پیشنهاد رزومه‌های مشابه به متن فرصت شغلی و نمایش آن‌ها به کارفرما.

۴.۶ محدودیت‌های تحقیق

در هر پژوهش آکادمیک یا در سطح دنیای واقعی که مبتنی بر داده هستند، محدودیت‌هایی وجود دارد. زیرا هیچ‌گاه نمی‌توان مطمئن شد داده‌های انسانی کافی هستند و به هر تعداد داده ذخیره شود، نمی‌توان به طور قطعی مدعی شد که داده‌های بعدی نیز تکرار داده‌های کنونی هستند. از این رو یکی از اصلی‌ترین محدودیت‌ها در این پژوهش کمبود داده است. داده‌ی رزومه، از آن نوع نیست که بتوان به راحتی آن را جمع آوری کرد یا به طور رایگان در اختیار داشت. این مجموعه داده، مخصوصا نوع فارسی آن توسط شرکت‌های بزرگ جمع‌آوری می‌شود و به دلیل در خطر افتادن اطلاعات شخصی افراد، نمی‌توان آن را به راحتی در اختیار داشت. در مورد داده‌ی انگلیسی هم این محدودیت وجود دارد اما می‌توان یک یا دو مجموعه دادگان در این حوزه را در اینترنت یافت. هم‌چنین یکی دیگر از نکات مهم موضوعات رزومه است. در حالی که شغل‌های موجود در دنیا بسیار متنوع هستند، اما مجموعه رزومه‌هایی که می‌توان یافت عموما در حوزه‌ی فناوری اطلاعات یا کارهای دفتری هستند و دشوار است که بتوان مجموعه‌ای رزومه با هر تعداد موضوع دلخواه را در اختیار داشت.

۵.۶ نتیجه‌گیری و کارهای آتی

تعدا بی‌شماری محدودیت در دنیای کسب و کار وجود دارد که مانع شرکت‌ها می‌شود تا پیشرفت کنند. یکی مهم‌ترین عوامل، کارمندان یک شرکت هستند. اگر کافرما کارمندان مناسبی برای شرکت استخدام نکند، این کارمندان از ابتدا به خوبی وظایف خود را خوب انجام نمی‌دهند یا ممکن است با وجود توانایی‌های بالا، در جایگاه پایین‌تری استخدام شوند و به زودی از کار خود خسته شوند. یکی از راه‌های حل این مشکل، استخدام افراد در جای مناسب است. از طرفی با وجود عمومیت یافتن شبکه‌های مجازی و امکان دیده شدن فرصت‌های شغلی توسط افراد بیشتر، تعداد اشخاصی که برای شغل‌های مختلف اقدام می‌کنند، روز به روز در حال افزایش است. در پی این گسترش، تعداد رزومه‌های ورودی شرکت‌ها و زمان صرف‌شده روی آن‌ها افزایش و دقت کاهش می‌یابد. بنابراین نیاز یک سیستم هوشمند برای خوانش و پردازش این رزومه‌ها اجتناب‌ناپذیر است.

در این پژوهش سیستمی طراحی شد که سعی بر حل کامل این مشکل دارد. این سیستم، از بین تعداد زیادی از رزومه،تعدادی از آن‌ها را به عنوان لیست کوتاه انتخاب کرده و به کارفرما نمایش می‌دهد.

در ادامه‌ی این راه، تصمیم داریم قسمت ترجمه‌ی رزومه‌های انگلیسی به فارسی را بهبود ببخشیم و سعی کنیم به جای یک فایل ترجمه‌ی کلمات، از یک سیستم مدیریت دیتابیس استفاده کنیم. هم‌چنین مجموعه‌ی دادگان خود را بزرگتر کنیم به گونه‌ای که بتوان مدل‌های شبکه‌ی عصبی را با استفاده از تنها متن‌های رزومه آموزش دهیم. هم‌چنین قصد داریم شبکه‌ی عصبی متناسب با این پژوهش را طراحی کنیم به جای آن‌که مدل‌های آماده را استفاده کنیم.

مراجع

[1] Roy, Pradeep Kumar et al. “A Machine Learning approach for automation of Resume Recommendation system.” *Procedia Computer Science* 167 (2020): 2318-2327.

[2] Delgado, Henrique José Trindade. “Applying Machine Learning Techniques to Implement a Decision Support Mechanism for Human Resources Recruitment.” (2019).

[3] Yasmin, Farzana & Nur, Mohammad Imtiaz & Shamsul, Mohammad. (2019). Developing a Framework for Potential Candidate Selection. International Journal of Advanced Computer Science and Applications. 10. 10.14569/IJACSA.2019.0101245.

[4] Catherine, Rose & Visweswariah, Karthik & Chenthamarakshan, Vijil & Kambhatla, Nanda. (2010). PROSPECT: A system for screening candidates for recruitment. 659-668. 10.1145/1871437.1871523.

[5] Elsafty, Ahmed & Riedl, Martin & Biemann, Chris. (2018). Document-based Recommender System for Job Postings using Dense Representations. 216-224. 10.18653/v1/N18-3027.

[6] Verma, Mayuri. “Cluster based Ranking Index for Enhancing Recruitment Process using Text Mining and Machine Learning.” *International Journal of Computer Applications* 157 (2017): 23-30.

[7] V, Senthil kumaran & Annamalai, Sankar. (2013). Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping EXPERT. International Journal of Metadata, Semantics and Ontologies. 8. 56-64. 10.1504/IJMSO.2013.054184.

[8]Palshikar, Girish Keshav et al. “Automatic Shortlisting of Candidates in Recruitment.” *ProfS/KG4IR/Data:Search@SIGIR* (2018).

[9] Faliagka, Evanthia & Ramantas, Kostas & Tsakalidis, Athanasios & Tzimas, Giannis. (2012). Application of Machine Learning Algorithms to an online Recruitment System.

[10] Rafter, Rachael & Bradley, Keith & Smyth, Barry. (2000). Personalised Retrieval for Online Recruitment Services.

[11] Carlos Del Cacho & Estrella Pulido, (2015) A Language Model-based Job Recommender.

[12] Thomas Schmitt, Francois Gonard, Phillipe Caillou, Michèle Sebag. Language Modelling for Collaborative Filtering: Application to Job Applicant Matching. In 29th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2017, Boston, MA, USA, November 6-8, 2017. pages 1226-1233, IEEE Computer Society, 2017.

[13] Maheshwary, Saket & Misra, Hemant. (2018). Matching Resumes to Jobs via Deep Siamese Network. 10.1145/3184558.3186942.

[14] Maheshwary, Saket & Misra, Hemant. (2018). Matching Resumes to Jobs via Deep Siamese Network. 10.1145/3184558.3186942.

[15] D. Mhamdi, R. Moulouki, M.Y. El Ghoumari, M. Azzouazi, L. Moussaid,

Job Recommendation based on Job Profile Clustering and Job Seeker Behavior,Procedia Computer science,Volume 175, 2020, Pages 695-699, ISSN 1877-0509, j.procs.2020.07.102.

[16] Jeevankrishna (2020). Job recommendation system using machine learning and natural language processing. Master's Thesis, Dublin Business School.

[17] Paparrizos, Ioannis & Cambazoglu, Berkant & Gionis, Aristides. (2011). Machine learned job recommendation. 325-328. 10.1145/2043932.2043994.

[18] Hong, Wenxing & Zheng, Siting & Wang, Huan. (2013). Dynamic user profile-based job recommender system. Proceedings of the 8th International Conference on Computer Science and Education, ICCSE 2013. 1499-1503. 10.1109/ICCSE.2013.6554164.

[19] Shivam Bansal, Aman Srivastava, Anuja Arora, Topic Modeling Driven Content Based Jobs Recommendation Engine for Recruitment Industry, Procedia Computer Science, Volume 122, 2017, Pages 865-872, ISSN 1877-0509, 10.1016/j.procs.2017.11.448.

[20] Guo, Shiqiang et al. "RésuMatcher: A personalized résumé-job matching system." *Expert Syst. Appl.* 60 (2016): 169-182.

[21] Bhatia, V., Rawat, P., Kumar, A., & Shah, R.R. (2019). End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT. ArXiv, abs/1910.03089.

[22] Bhola, Akshay & Halder, Kishaloy & Prasad, Animesh & Kan, Min-Yen. (2020). Retrieving Skills from Job Descriptions: A Language Model Based Extreme Multi-Label Classification Framework. 5832-5842. 10.18653/v1/2020.coling-main.513.

[23] Valverde-Rebaza, Jorge & Puma, Ricardo & Bustios, Paul & Silva, Nathalia. (2018). Job Recommendation based on Job Seeker Skills: An Empirical Study.

[24] Alghieth, Manal & Al-Shargabi, Amal. (2019). A Map-based Job Recommender Model. International Journal of Advanced Computer Science and Applications. 10. 345-351. 10.14569/IJACSA.2019.0100945.

[25] Alotaibi, Shaha. (2012). A survey of job recommender systems. International Journal of the Physical Sciences. 7. 10.5897/IJPS12.482.

[26] Chen, Jie et al. "A Two-Step Resume Information Extraction Algorithm." *Mathematical Problems in Engineering* 2018 (2018): 1-8.

[27] Zu, Shicheng and Xiulai Wang. "Resume Information Extraction with A Novel Text Block Segmentation Algorithm." *International Journal on Natural Language Computing* (2019): n. pag.

[28] Ling, Wang & Dyer, Chris & Black, Alan & Trancoso, Isabel. (2015). Two/Too Simple Adaptations of Word2Vec for Syntax Problems. 10.3115/v1/N15-1142.