دانشکده مهندسی کامپیوتر

دستیار آزمایشگاه

آزمایشگاه داده کاوی دانشگاه علم و صنعت

دانیال کمالی

نام استاد کارآموزی:

دکتر صالح اعتمادی

مهر ماه ۱۳۹۸


بسم الله الرحمن الرحیم

تأییدیه‌ی صحت و اصالت نتایج

بسمه تعالی

اینجانب دانیال کمالی به شماره دانشجویی ۹۵۵۲۱۳۹۶ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی مطالب مندرج در این گزارش حاصل ۳۰۰ ساعت حضور و کار اینجانب در آزمابشگاه داده کاوی دانشگاه علم و صنعت و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری‌شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم آموزشی، پژوهشی و انضباطی با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم.

نام و نام خانوادگی:

امضا و تاریخ:

تشکر و قدردانی:

با تشکر از جناب آقای جان فدا که در تمامی مراحل تحقیقات به بنده کمک و یاری رساندند.

چکیده
آزمایشگاه داده کاوی با تمرکز بر روی پردازش متن طبیعی در درحال توسعه یک گراف دانش برای زبان فارسی است، در طی این دوره کارآموزی بنده بر روی ویژگی های مختلف زبان فارسی تحقیق انجام دادم که حاصل آن یک تحقیق در موضوع معرفه و نکره در زبان فارسی و یک مقاله در حوزه ابزار های Tokenization در زبان فارسی است.

واژه‌های کلیدی: داده کاوی ،پردازش متن طبیعی،Tokenization،معرفه و نکره

فصل ۱:

معرفی حوزه کارآموزی

آزمایشـــگاه داده کاوی تحت نظر دکتر مینایی و مســـئولیت دکتر حســـینی با تمرکز بر روی پردازش متن طبیعی در درحال تو سعه یک گراف دانش برای زبان فار سی ا ست، در این آزمایـ شگاه چندین گروه زیر نظر چندین ناظر مشـغول انجام فعالیت هسـتند و هرگروه درحال توسـعه ی بخشـی از گراف دانش اسـت.همه تیم ها در روزی مشـخص جلسـه اسـکرام برگزار میکنند و طی آن درباره کار های محوله هفته قبل توضـیح ارائه میدهند و تسک های جدید یا راهنمایی حل مشکل به اعضا داده میشود.

فصل ۲:

مشروح فعالیت های انجام شده در محل استقرار

در ابتدا لیستی از تکنولوژی های در حال استفاده در گروه ها و موضوعات و مراجع لازم به بنده داده شد تا درباره موضوعات مطالعه کنم و حوزه مورد علاقه خودم را به اطلاع برسانم تا در انتخاب تیم بنده مورد توجه قرار دهند.

بعد از مشخص شدن تیم با توجه به نیاز سیستم گراف دانش به تشخیص معرفه و نکره به بنده یک تحقیق حول معرفه و نکره در زبان فارسی محول شد،درمرحله بعد با توجه به مشکلات Tokenizer فارسی موجود در ابتدا به بنده یک تحقیق درباره انواع Tokenizer های فارسی داده شد. پس از ارائه گزارش، بنده مسئولیت بررسی و مقایسه هر کدام رو براساس یک پیکره ی فارسی را به عهده گرفتم که به صورت یک مقاله به انجام رسید.

فصل ۳:

نتیجه گیری و پیشنهادها

کسی که میخواهد برای کارآموزی در این آزمایشگاه اقدام کند بهتر است قبل از شروع درباره آزمایشگاه تحقیق کند و در صورت علاقه به موضوع و داشتن دانش قبلی درباره موضوعات پردازش متن طبیعی به کار در آزمایشگاه اقدام کند.