درس NLP
دکتر داوودآبادی

کارگاه آشنایی با ابزار های هضم و حرف
امیررضا ستارزاده

(1) روند شروع به کار یک محصول ASR* و توسعه آن را با توجه به نکات گفته شده سر کارگاه ، توضیح دهید.

(2) فرض کنید چند مدل (برای مثال whisper – harf – mozilla – google – nemo– ....) برای تبدیل صوت به متن فارسی در دسترس داریم و هرکدام 16 خروجی با ترتیب اولویت احتمال درستی میدهد ، برای اینکه از بین این تعداد زیاد خروجی ، بتوانیم تعداد 16 خروجی نهایی انتخاب کنیم و خروجی دهیم حداقل دو روش(که در عمل استفاده میشود یا اگر روش جدیدی به ذهنتنان میرسد که منطقی است)توضیح دهید.

(3) مفهوم توابع normalizer , formalizer, lemmatizer, stemmer, chunker, tagger, postagger, embedder, wordembedder, parser را در دنیای پردازش متن توضیح دهید.

(4) ویدیوی Turing را به مدل حرف دادیم و این خروجی را به ما داد. فایل زیرنویس منتشر شده از سمت خود سازنده ویدیو را که در این لینک قابل دانلود است ، با آن تطابق دهید و پس از نرمالسازی ، WER* , CER* خروجی مدل را با استفاده از کتابخانه هضم و لینک1 و لینک2 حساب کنید.

(5) تعداد فعل ها و قیدهای موجود در دوفایل مورد بحث در سوال قبل را با استفاده از کتابخانه هضم محاسبه کنید.

(6) ریشه فعلی که در هر فایل بیشترین تکرار را داشته است با استفاده از کتابخانه هضم بیابید.

(7) با استفاده از هضم ، کدی بنویسید که لیستی از لیست های حاوی 5 کلمه فارسی بگیرد و کله ی بی ربط نسبت به بقیه لیست درهر گروه را برگرداند.

**** asr = automatic speech recognition
**** wer = word error rate
**** cer = character error rate