دانشکده مهندسی کامپیوتر

پردازش زبان‌های طبیعی

دکتر بهروز مینایی

تمرین سری اول **زمان تحویل:** ۱۱ آبان

۱- متن داستان آلیس در سرزمین عجایب در ضمیمه موجود است.
در این داستان در کل چند کلمه وجود دارد؟ چند کلمه متفاوت وجود دارد؟ آیا رابطه Church and Gale برقرار است؟ $|V| > O(N^{½})$

۲- ریشه تمام کلمات موجود در داستان آلیس در سرزمین عجایب را به‌وسیله الگوریتم Porter استخراج‌کنید. چند ریشه متفاوت در این داستان وجود دارد؟ لیست آن‌ها را ارسال کنید.

۳- یک فایل متنی ساده (با تعداد کلمات کافی) مانند داستان آلیس در سرزمین عجایب را پردازش کرده و یک مدل Bigram از آن بسازید. سپس تعدادی جمله ممکن را بر اساسBigram های موجود بر اساس بازی Shannon تولید کنید. این کار را با Trigram نیز انجام دهید. می‌توانید این کار را با نرم‌افزارهای موجود انجام دهید.

۴- در مسئله‌ی ماهی‌ها (در بخش Good Turing intuition) احتمال یافتن همه‌ی ماهی‌ها در ماهیگیری بعدی را با روش Good-Turing محاسبه نمایید. در مورد مجموع احتمال آن‌ها تحقیق کنید.

در هر سؤال گزارش خود را در قالب یک فایل pdf آماده کرده و کدها و پیکره‌های تولیدشده خود را همراه با گزارش در یک فایل zip در سایت بارگذاری کنید.

در صورت وجود مشکل، سؤالات خود را به ایمیل‌های majid.asgari@gmail.com و y.alizadeh@yahoo.com ارسال کنید.