

# پردازش زبان طبیعی با پایتون

مؤلفین

هابسون لین

کول هاوارد

هانس ماکس هاپکه

مترجم

ایوب ترکیان

نیاز دانش



## فهرست مطالب

شماره صفحه	عنوان
۹	فصل ۱ / مرور کلی NLP
۹	۱.۱ زبان طبیعی و زبان برنامه‌سازی
۱۰	۲.۱ جادو
۱۱	۱.۲.۱ ماشین سخنگو
۱۲	۲.۲.۱ ریاضی
۱۴	۳.۱ کاربردهای عملی
۱۶	۴.۱ زبان از نگاه رایانه
۱۷	۱.۴.۱ زبان قفل‌ها
۱۸	۲.۴.۱ عبارات منظم
۲۰	۳.۴.۱ روایات سخنگوی ساده
۲۴	۴.۴.۱ روش دیگر
۲۹	۵.۱ فرافضا
۳۱	۶.۱ ترتیب واژه و گرامر
۳۲	۷.۱ خط‌مسیر زبان طبیعی روایات گفتگو
۳۵	۸.۱ پردازش در عمق (عمیق)
۳۷	۹.۱ IQ زبان طبیعی
۴۰	۱۰.۱ خلاصه
۴۱	فصل ۲ / توکن‌سازی واژه
۴۳	۱.۲ چالش‌ها (پیش‌نمایش ریشه‌یابی)
۴۴	۲.۲ ساخت فرهنگ لغات با توکن‌ساز
۵۴	۱.۲.۲ ضرب داخلی
۵۵	۲.۲.۲ سنجش هم‌پوشانی کیفیت واژه
۵۶	۳.۲.۲ بهبود توکن
۶۲	۴.۲.۲ بسط فرهنگ‌نامه با $n$ -گرام‌ها
۶۸	۵.۲.۲ نرمال‌سازی قاموس
۷۶	۳.۲ تمایل
۷۹	۱.۳.۲ VADER - تحلیل‌گر تمایل قاعده‌محور
۸۰	۲.۳.۲ بیز ساده‌انگاران
۸۴	۴.۲ خلاصه
۸۵	فصل ۳ / ریاضی کلمات (بردارهای TF-IDF)
۸۶	۱.۳ کیفیت واژه
۹۱	۲.۳ بردارسازی
۹۴	۱.۲.۳ فضاهای برداری
۹۹	۳.۳ قانون زیف
۱۰۲	۴.۳ مدل‌سازی موضوعی
۱۰۵	۱.۴.۳ بازگشت به زیف
۱۰۶	۲.۴.۳ درجه‌بندی مربوط بودن

۱۰۹	ابزار	۳.۴.۳
۱۱۰	آلترناتیوها	۴.۴.۳
۱۱۰	okapi BM25	۵.۴.۳
۱۱۲	مرحله بعدی	۶.۴.۳
۱۱۲	خلاصه	۵.۳

#### فصل ۴ / تحلیل معنایی

۱۱۴	تعداد کلمه و نمرات موضوع	۱.۴
۱۱۴	بردارهای TF-IDF و لم‌سازی	۱.۱.۴
۱۱۶	بردارهای موضوعی	۲.۱.۴
۱۱۸	آزمایش ذهنی	۳.۱.۴
۱۲۲	الگوریتم رتبه‌بندی موضوعی	۴.۱.۴
۱۲۴	طبقه‌گر LDA	۵.۱.۴
۱۲۹	تحلیل سمانتیک نهفته	۲.۴
۱۳۲	آزمایش ذهنی واقعی	۱.۲.۴
۱۳۴	تجزیه مقدار ویژه	۳.۴
۱۳۶	U- بردارهای ویژه چپ	۱.۳.۴
۱۳۷	S- مقادیر ویژه	۲.۳.۴
۱۳۸	VT- بردارهای ویژه راست	۳.۳.۴
۱۳۹	جهت ماتریس SVD	۴.۳.۴
۱۴۰	کوتاه‌سازی موضوعات	۵.۳.۴
۱۴۲	آنالیز مؤلفه اصلی	۴.۴
۱۴۴	PCA روی بردارهای ۳-بُعدی	۱.۴.۴
۱۴۵	بازگشت به NLP	۲.۴.۴
۱۴۸	استفاده PCA در تحلیل معنایی SMS	۳.۴.۴
۱۵۰	SVD کوتاه‌شده تحلیل معنایی	۴.۴.۴
۱۵۱	دسته‌بندی اسپم با LSA	۵.۴.۴
۱۵۴	تخصیص دیریکله نهفته (LDiA)	۵.۴
۱۵۵	ایده LDiA	۱.۵.۴
۱۵۷	مدل موضوعی LDiA پیام کوتاه	۲.۵.۴
۱۶۰	LDA+LDiA = طبقه‌گر اسپم	۳.۵.۴
۱۶۲	مقایسه منصفانه: ۳۲ موضوع LDiA	۴.۵.۴
۱۶۴	فاصله و مشابهت	۶.۴
۱۶۷	حرکت با پس‌خور	۷.۴
۱۶۸	تحلیل تمایز خطی	۱.۷.۴
۱۷۰	قدرت بردار موضوعی	۸.۴
۱۷۰	جستجوی سمانتیک	۱.۸.۴
۱۷۳	بهبودها	۲.۸.۴
۱۷۳	خلاصه	۹.۴

#### فصل ۵ / مبانی شبکه‌های عصبی

۱۷۶	شبکه‌های عصبی	۱.۵
۱۷۶	پرسپترون	۱.۱.۵
۱۷۷	پرسپترون عددی	۲.۱.۵
۱۷۸	انحرافی به بایاس	۳.۱.۵

۱۹۳	.....	سطح خطا	۴.۱.۵
۱۹۴	.....	شیب	۵.۱.۵
۱۹۵	.....	اصلاح رویه	۶.۱.۵
۱۹۷	.....	کراس: شبکه‌های عصبی در پایتون	۷.۱.۵
۲۰۰	.....	شبکه‌های عمیق	۸.۱.۵
۲۰۰	.....	نرمال‌سازی	۹.۱.۵
۲۰۱	.....	خلاصه	۲.۵

## فصل ۶ / استدلال با بردارهای واژه

۲۰۴	.....	۱.۶	پرسمان‌های سمانتیک و تشبیه
۲۰۴	.....	۱.۱.۶	سوالات تشبیهی
۲۰۶	.....	۲.۶	بردارهای واژه
۲۱۰	.....	۱.۲.۶	استدلال با گرایش برداری
۲۱۳	.....	۲.۲.۶	محاسبه نمایش‌های Word2vec
۲۲۳	.....	۳.۲.۶	ماژول genism.word2vec
۲۲۵	.....	۴.۲.۶	تولید نمایش بردار واژه
۲۲۸	.....	۵.۲.۶	Word2vec و GloVe
۲۲۹	.....	۶.۲.۶	fastText
۲۲۹	.....	۷.۲.۶	Word2vec و LSA
۲۳۱	.....	۸.۲.۶	مصورسازی روابط واژه
۲۳۱	.....	۹.۲.۶	واژه‌های غیرطبیعی
۲۳۱	.....	۱۰.۲.۶	مشابهت سند با Doc2vec
۲۴۱	.....	۳.۶	خلاصه

## فصل ۷ / شبکه عصبی کانولوشن

۲۴۴	.....	۱.۷	یادگیری معنا
۲۴۶	.....	۲.۷	جعبه‌ابزار
۲۴۷	.....	۳.۷	شبکه عصبی کانولوشن
۲۴۷	.....	۱.۳.۷	بلوک‌های ساخت
۲۵۰	.....	۲.۳.۷	اندازه گام (پرش)
۲۵۰	.....	۳.۳.۷	ترکیب فیلتر
۲۵۲	.....	۴.۳.۷	لایه‌گذاری
۲۵۴	.....	۵.۳.۷	یادگیری
۲۵۵	.....	۴.۷	پنجره‌های کم‌عرض
۲۵۶	.....	۱.۴.۷	پیاده‌سازی در کراس: آماده‌سازی داده‌ها
۲۶۲	.....	۲.۴.۷	معماری شبکه عصبی کانولوشن
۲۶۳	.....	۳.۴.۷	رای‌گیری
۲۶۶	.....	۴.۴.۷	دورریزی
۲۶۷	.....	۵.۴.۷	گیلاس روی بستنی
۲۶۹	.....	۶.۴.۷	یادگیری (آموزش)
۲۷۱	.....	۷.۴.۷	استفاده از مدل در خط‌مسیر
۲۷۳	.....	۸.۴.۷	گام بعدی
۲۷۴	.....	۵.۷	خلاصه

۲۷۵	فصل ۸ / شبکه عصبی برگشتی
۲۷۸	۱.۸ خاطر آوری با شبکه‌های برگشتی
۲۸۳	۱.۱.۸ پس انتشار در زمان
۲۸۶	۲.۱.۸ به روزرسانی
۲۸۸	۳.۱.۸ خلاصه
۲۸۹	۴.۱.۸ هزینه محاسباتی
۲۸۹	۵.۱.۸ شبکه عصبی برگشتی با کراس
۲۹۴	۲.۸ سر هم کردن موارد
۲۹۶	۳.۸ یادگیری
۲۹۷	۴.۸ فرآپارامترها
۳۰۰	۵.۸ پیش بینی کردن
۳۰۱	۱.۵.۸ حالت دار بودن
۳۰۲	۲.۵.۸ جاده دوطرفه
۳۰۴	۳.۵.۸ این چیست؟
۳۰۴	۶.۸ خلاصه

۳۰۵	فصل ۹ / شبکه عصبی LSTM
۳۰۶	۱.۹ LSTM
۳۱۶	۱.۱.۹ پس انتشار در زمان
۳۱۹	۲.۱.۹ پیش بینی
۳۲۰	۳.۱.۹ داده‌های غیر تمیز
۳۲۳	۴.۱.۹ بازگشت به داده‌های غیر تمیز
۳۲۵	۵.۱.۹ کلمات و حروف
۳۳۱	۶.۱.۹ چت کردن
۳۳۴	۷.۱.۹ گفتار واضح
۳۴۱	۸.۱.۹ محتوای گفتار
۳۴۲	۹.۱.۹ انواع دیگر حافظه
۳۴۲	۱۰.۱.۹ تعمیق
۳۴۴	۲.۹ خلاصه

۳۴۵	فصل ۱۰ / مدل‌های توالی به توالی و توجه
۳۴۵	۱.۱۰ معماری رمزگذار-رمزدا
۳۴۷	۱.۱.۱۰ رمزدایی فکر
۳۵۰	۲.۱.۱۰ مشابهت با خود رمزگذار
۳۵۱	۳.۱.۱۰ گفتگوی توالی به توالی
۳۵۱	۴.۱.۱۰ مرور LSTM
۳۵۳	۲.۱۰ سرهم سازی خط مسیر توالی به توالی
۳۵۳	۲.۱.۱۰ آماده سازی برای آموزش
۳۵۴	۲.۲.۱۰ مدل توالی به توالی در کراس
۳۵۵	۳.۲.۱۰ رمزگذار توالی
۳۵۶	۴.۲.۱۰ رمزدای فکر
۳۵۸	۵.۲.۱۰ سرهم سازی شبکه توالی به توالی
۳۵۹	۳.۱۰ آموزش شبکه توالی به توالی
۳۶۰	۱.۳.۱۰ تولید توالی‌های خروجی

۳۶۱	ساخت بات چت	۴.۱۰
۳۶۲	آماده‌سازی واژگان برای آموزش	۱.۴.۱۰
۳۶۳	ساخت واژه‌نامه کاراکتر	۲.۴.۱۰
۳۶۳	تولید مجموعه‌های آموزش تک‌نمادی	۳.۴.۱۰
۳۶۴	آموزش بات چت	۴.۴.۱۰
۳۶۵	سرهم‌سازی مدل برای تولید توالی	۵.۴.۱۰
۳۶۵	پیش‌بینی توالی	۶.۴.۱۰
۳۶۶	تولید پاسخ	۷.۴.۱۰
۳۶۷	گفتگو با بات چت	۸.۴.۱۰
۳۶۷	بهیودها	۵.۱۰
۳۶۸	کاهش پیچیدگی آموزش	۱.۵.۱۰
۳۶۹	توجه کردن	۲.۵.۱۰
۳۷۰	در دنیای واقعی	۶.۱۰
۳۷۲	خلاصه	۷.۱۰

۳۷۳	فصل ۱۱ / استخراج اطلاعات	
۳۷۳	هستارهای نام‌دار و روابط	۱.۱۱
۳۷۴	پایگاه دانش	۱.۱.۱۱
۳۷۷	استخراج اطلاعات	۲.۱.۱۱
۳۷۸	الگوهای منظم	۲.۱۱
۳۷۹	عبارات منظم	۱.۲.۱۱
۳۸۰	استخراج ویژگی ML	۲.۲.۱۱
۳۸۲	اطلاعات با ارزش استخراج	۳.۱۱
۳۸۲	استخراج محل‌های GIS	۱.۳.۱۱
۳۸۳	استخراج تاریخ	۲.۳.۱۱
۳۸۸	استخراج ارتباطات (روابط)	۴.۱۱
۳۸۹	برچسب‌زنی ادات سخن (POS)	۱.۴.۱۱
۳۹۳	نرمال‌سازی هستار نام‌دار	۲.۴.۱۱
۳۹۴	نرمال‌سازی و استخراج ارتباط	۳.۴.۱۱
۳۹۵	الگوهای کلمه	۴.۴.۱۱
۳۹۵	بخش‌بندی	۵.۴.۱۱
۳۹۷	کار نکردن تقسیم («!؟»)	۶.۴.۱۱
۳۹۸	بخش‌بندی جمله با عبارت منظم	۷.۴.۱۱
۴۰۰	در دنیای واقعی	۵.۱۱
۴۰۱	خلاصه	۶.۱۱

۴۰۳	فصل ۱۲ / موتورهای دیالوگ	
۴۰۴	مهارت زبان	۱.۱۲
۴۰۵	رویکردهای جدید	۱.۱.۱۲
۴۱۱	رویکرد هیبریدی	۲.۱.۱۲
۴۱۱	رویکرد انطباق الگو	۲.۱۲
۴۱۳	بات چت انطباق الگو با AIML	۱.۲.۱۲
۴۲۱	نمای شبکه‌ای انطباق الگو	۲.۲.۱۲
۴۲۱	استقرار	۳.۱۲
۴۲۳	بازیابی (جستجو)	۴.۱۲

۴۲۴	چالش سیاق	۱.۴.۱۲
۴۲۶	نمونه بات چت بازیابی پایه	۲.۴.۱۲
۴۳۰	بات چت جستجوی پایه	۳.۴.۱۲
۴۳۳	مدل‌های زایشی (مولد)	۵.۱۲
۴۳۳	چت در مورد nlpia	۱.۵.۱۲
۴۳۵	محاسن و معایب دو رویکرد	۲.۵.۱۲
۴۳۵	ماشین قدرتی	۶.۱۲
۴۳۶	Will	۱.۶.۱۲
۴۳۷	فرایند طراحی	۷.۱۲
۴۴۰	شگرد	۸.۱۲
۴۴۰	سوالات با پاسخ قابل پیش‌بینی	۱۸.۱۲
۴۴۱	جذاب بودن	۲۸.۱۲
۴۴۲	جستجو	۳۸.۱۲
۴۴۲	اقبال	۴۸.۱۲
۴۴۲	وصل یا فصل	۵۸.۱۲
۴۴۲	احساسی شدن	۶۸.۱۲
۴۴۳	در دنیای واقعی	۹.۱۲
۴۴۴	خلاصه	۱۰.۱۲

### فصل ۱۳ / فرامقیاس کردن

۴۴۵	داده‌ها	۱.۱۳
۴۴۶	بهینه‌سازی الگوریتم‌های NLP	۲.۱۳
۴۴۷	نمایه‌سازی	۱.۲.۱۳
۴۴۸	نمایه‌سازی پیشرفته	۲.۲.۱۳
۴۵۰	نمایه‌سازی پیشرفته با Annoy	۳.۲.۱۳
۴۵۵	کاربرد نمایه‌های تقریبی	۴.۲.۱۳
۴۵۶	نایبوسته‌سازی برای بای‌پس نمایه	۵.۲.۱۳
۴۵۷	الگوریتم‌های RAM ثابت	۳.۱۳
۴۵۷	Gensim	۱.۳.۱۳
۴۵۸	محاسبه گراف	۲.۳.۱۳
۴۵۹	موازی‌سازی محاسبات NLP	۴.۱۳
۴۵۹	آموزش مدل‌های NLP روی GPU	۱.۴.۱۳
۴۶۱	اجاره یا خرید	۲.۴.۱۳
۴۶۱	گزینه‌های اجاره GPU	۳.۴.۱۳
۴۶۲	واحدهای پردازش تانسور	۴.۴.۱۳
۴۶۲	کاهش جایای حافظه در زمان آموزش مدل	۵.۱۳
۴۶۶	اشراف به مدل با TensorBoard	۶.۱۳
۴۶۶	مصورسازی نهفته‌سازی کلمه	۱.۶.۱۳
۴۶۸	خلاصه	۷.۱۳
۴۶۹	پیوست الف/ ابزار NLP	
۴۷۷	پیوست ب/ پایتون سرگرمی و عبارات منظم	
۴۸۳	پیوست ج/ بردارها و ماتریس‌ها	
۴۸۹	پیوست د/ ابزارها و تکنیک‌های یادگیری ماشین	
۵۰۳	پیوست ه/ برپایش AWS GPU	
۵۱۷	پیوست و/ هش‌سازی حساس به محلیت	

## فصل ۱

# مرور کلی NLP

شما در حال ورود به رویداد جالب پردازش زبان طبیعی (NLP) هستید. ابتدا، نشان داده خواهد شد NLP چیست و چه کارهایی با آن می‌توان انجام داد. این کار، چرخ‌ها را به حرکت در آورده، و به شما کمک می‌کند روش‌های استفاده از NLP در زندگی روزمره خویش، سر کار و در منزل، را بررسی کنید.

سپس، جزئیات دقیق نحوه پردازش مقدار کمی متن انگلیسی با استفاده از زبان برنامه‌سازی مثل پایتون بررسی شده، که کمک خواهد کرد به تدریج جعبه ابزار NLP خویش را درست کنید. در این فصل، اولین برنامه خویش را خواهید نوشت که می‌تواند جملات انگلیسی را خوانده و بنویسد. این قطعه برنامه پایتون یکی از تعداد زیاد برنامه‌هایی است که استفاده خواهید کرد، تا همه شگردهای مورد نیاز برای سرهم‌سازی یک موتور دیالوگ زبان انگلیسی، ماشین سخنگو، را یاد بگیرید.

## ۱.۱ زبان طبیعی و زبان برنامه‌سازی

زبان‌های طبیعی از زبان‌های برنامه‌سازی رایانه متفاوت هستند. قرار نیست آنها، مثل زبان‌های برنامه‌سازی، به یک مجموعه معین عملیات ریاضی، ترجمه شوند. زبان‌های طبیعی، مورد استفاده انسان‌ها برای به اشتراک‌گذاری اطلاعات با یکدیگر هستند. از زبان‌های برنامه‌سازی برای صحبت با یکدیگر راجع به وقایع روزانه یا برای آدرس مغازه خواروبارفروشی دادن، استفاده نمی‌شود. یک برنامه کامپیوتر نوشته شده با زبان برنامه‌سازی، به ماشین می‌گوید دقیقاً چه کاری را انجام بدهد. ولی برای زبان‌های طبیعی نظیر فارسی و عربی، کامپایلر یا مفسر وجود ندارد.



**تعریف** پردازش زبان طبیعی، یک زمینه تحقیق در علم کامپیوتر و هوش مصنوعی (AI) بوده که مربوط به پردازش زبان طبیعی نظیر انگلیسی یا ماندارین است. این پردازش به طور کلی شامل ترجمه زبان طبیعی به داده‌ها (اعداد) بوده که کامپیوتر بتواند برای یادگیری در مورد پدیده‌ها استفاده کند. و این شناخت جهان، در بعضی مواقع برای تولید متن زبان طبیعی استفاده شده، که انعکاس آن شناخت است.

با این حال، در این فصل نحوه پردازش زبان طبیعی توسط ماشین نشان داده می‌شود. ممکن است حتی این را یک مفسر زبان طبیعی، درست مشابه مفسر پایتون، در نظر بگیرید. در موقعی که برنامه کامپیوتری توسعه داده شده، زبان طبیعی را پردازش می‌کند، قادر به عمل روی آن عبارات یا حتی پاسخ به آنها خواهد بود. ولی این اقدامات و پاسخ‌ها دقیقاً تعریف نشده‌اند، که فضای اعمال سلیقه شما، توسعه‌گر خط مسیر زبان طبیعی، را فراهم می‌کند.

**تعریف** یک سیستم پردازش زبان طبیعی غالباً خط‌میسر نامیده شده، چون معمولاً شامل چند مرحله پردازش بوده که در آنها زبان طبیعی از یک طرف وارد شده، و خروجی پردازش شده از طرف دیگر به بیرون می‌رود.

به زودی، توان نوشتن نرم‌افزاری را در خود ایجاد خواهید کرد که کارهای جالب غیرقابل پیش‌بینی، نظیر انجام گفتگو را طوری صورت داده که به نظر برسد ماشین‌ها یک مقدار رفتار انسانی دارند. در ابتدا، ممکن است کاری که تکنولوژی پیشرفته انجام می‌دهد، یک مقدار سحرآمیز جلوه کند. ولی پرده که عقب زده شود، پشت صحنه را می‌توان بررسی کرد، و زود همه ابزار مورد نیاز برای انجام شگردهای جادویی را خود شما کشف خواهید کرد.

## ۲.۱ جادو

چه چیزی در باره اینکه یک ماشین بتواند در یک زبان طبیعی بخواند و بنویسد، جادویی است؟ ماشین‌ها از زمان اختراع کامپیوتر، زبان‌ها را پردازش می‌کرده‌اند. با این حال، این زبان‌های «رسمی»- نظیر زبان‌های اولیه Ada، COBOL، و فورترن، طوری طراحی می‌شدند که فقط به یک شیوه صحیح تفسیر (یا کامپایل) شوند. امروزه، ویکی‌پدیا لیستی از بیش از ۷۰۰ زبان برنامه‌سازی را فراهم می‌کند. بر خلاف این، ۱۰ برابر این مقدار زبان طبیعی محاوره‌ای در جهان وجود دارد. نمایه گوگل اسناد زبان طبیعی بیش از ۱۰۰ میلیون گیگابایت است. و این فقط نمایه است. و ناکامل است. اندازه محتوای زبان طبیعی واقعی برخط، در حال حاضر متجاوز از ۱۰۰