



گروه هوش مصنوعی، دانشکده مهندسی
کامپیوتر

بہ نام خدا



بخش اول

مقدمات پردازش زبان طبیعی

حمیدرضا برادران کاشانی





تعریف NLP

❖ زبان طبیعی زبانی است که در تعاملات اجتماعی روزمره با استفاده از آن می نویسیم و صحبت می کنیم.

❖ به بیان ساده منظور از زبان طبیعی، زبان انسانی است.

❖ پردازش زبان طبیعی (NLP) یکی از شاخه های علوم کامپیوتر و هوش مصنوعی است که به تعاملات بین کامپیوتر و انسان، از طریق زبان طبیعی می پردازد.

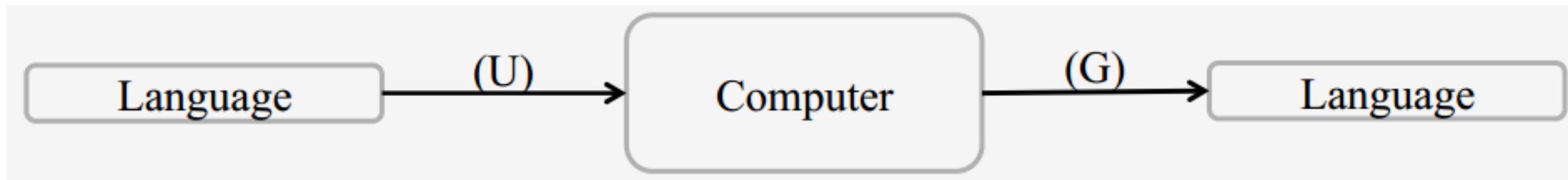
❖ به تعریف دقیق تر، پردازش زبان های طبیعی عبارت است از استفاده از کامپیوتر برای پردازش:
زبان گفتاری و زبان نوشتاری.

❖ بدین معنی که کامپیوترها را قادر سازیم که گفتار یا نوشتار تولید شده در قالب و ساختار یک زبان طبیعی را تحلیل و درک نموده یا آن را **تولید** نمایند.



تعریف NLP

- ❖ به عبارتی در زبان طبیعی بایستی کامپیوترها یاد بگیرند:
- ❖ چگونه می توان زبان را درک کرد (Language Understanding)?
- ❖ چگونه می توان زبان را تولید کرد (Language Generation)?



Hamidreza Baradaran Kashani



ارتباطات و زبان (Language and Communication)

❖ Speaker

- Intention (goals, shared knowledge and beliefs)
- Generation (tactical)
- Synthesis (text or speech)

❖ Listener

- Perception
- Interpretation (syntactic, semantic, pragmatic)
- Incorporation (internalization, understanding)

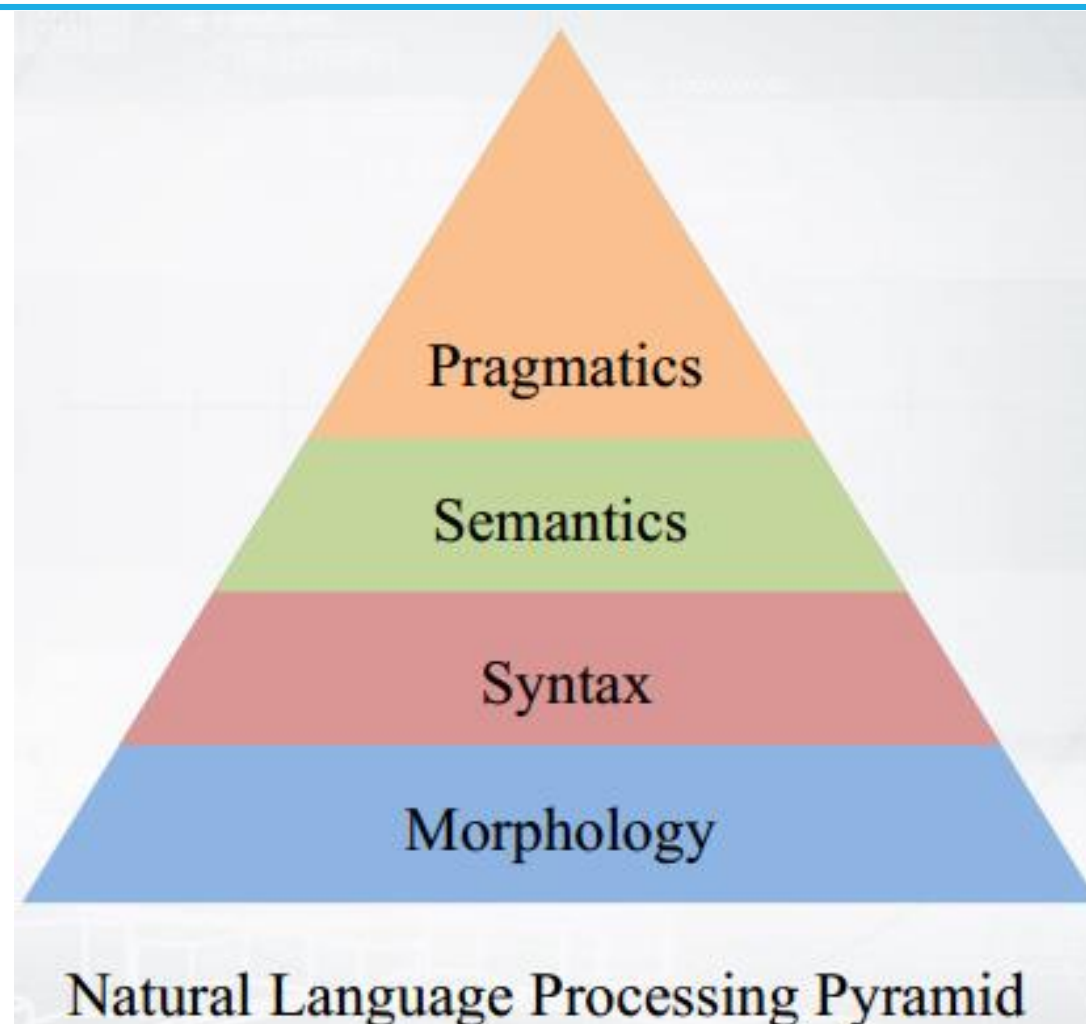
❖ Both

- Context (grounding)

Hamidreza Baradaran Kashani



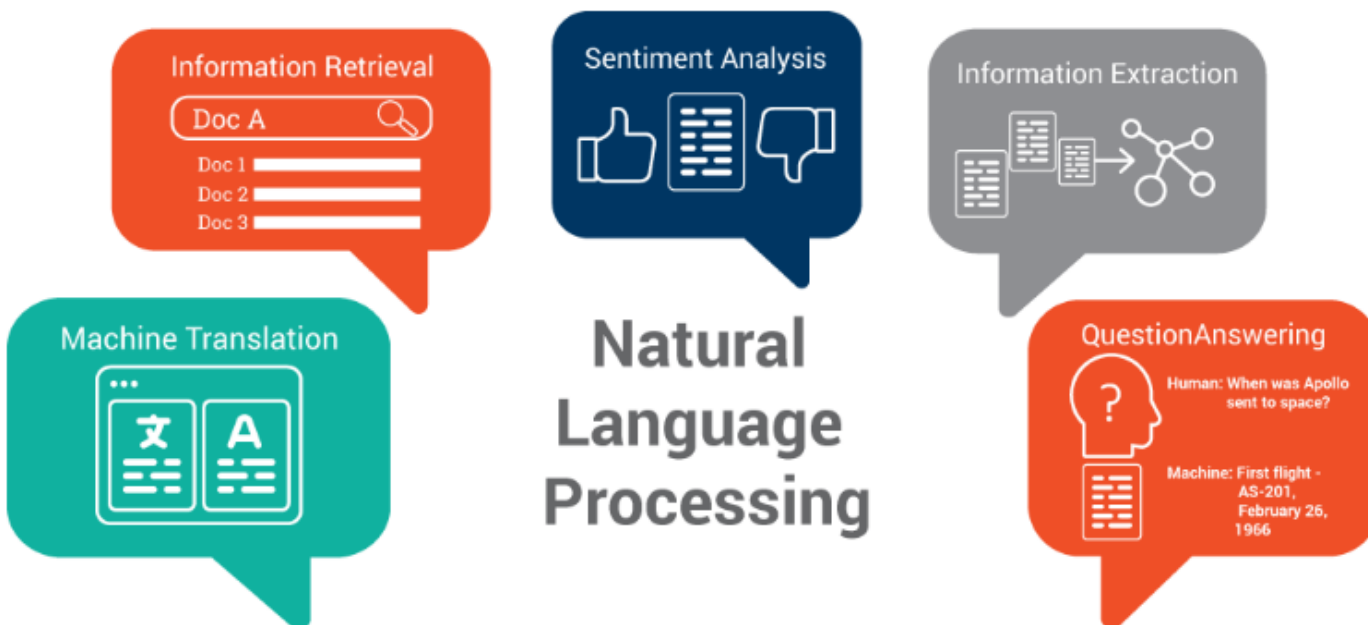
NLP Pyramid



Hamidreza Baradaran Kashani



کاربردهای پردازش زبان طبیعی؟



Hamidreza Baradaran Kashani



کاربردهای NLP

❖ فرض کنید می خواهیم بدانیم نقل قول زیر از کجا آمده است؟

Dave Bowman: Open the pod bay doors, HAL.
HAL: I'm sorry Dave. I'm afraid I can't do that.

❖ این نقل قول مربوط به یک فیلم با عنوان **2001:Space Odyssey** ساخته **Stanley Kubrick** در ۱۹۶۸ است.



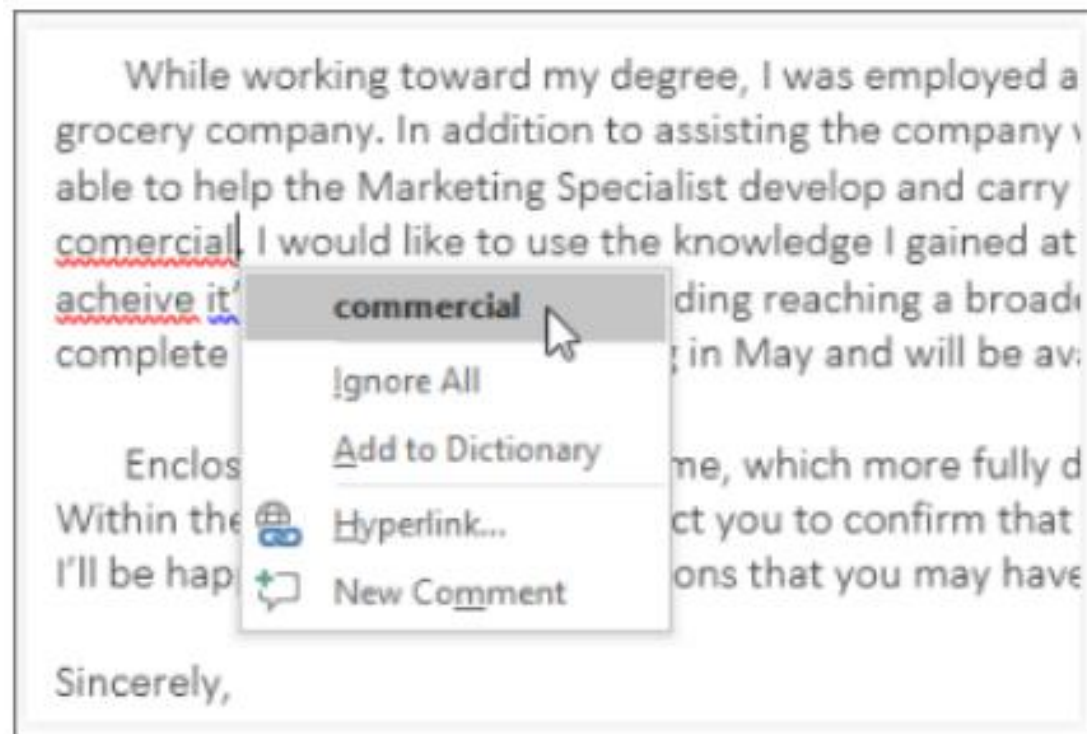
❖ این فیلم یکی از اولین داستان های علمی است که در آن کامپیوترها با انسان به زبان طبیعی تعامل می کنند.

Hamidreza Baradaran Kashani



کاربردهای NLP

❖ تصحیح خطاهای املائی (Spell Correction)



Hamidreza Baradaran Kashani



کاربردهای NLP

❖ استخراج اطلاعات (Information Extraction)

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jurafsky

Hi Dan, we've now scheduled the curriculum meeting.

It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

Create new Calendar entry



Event: Curriculum mtg

Date: Jan-16-2012

Start: 10:00am

End: 11:30am

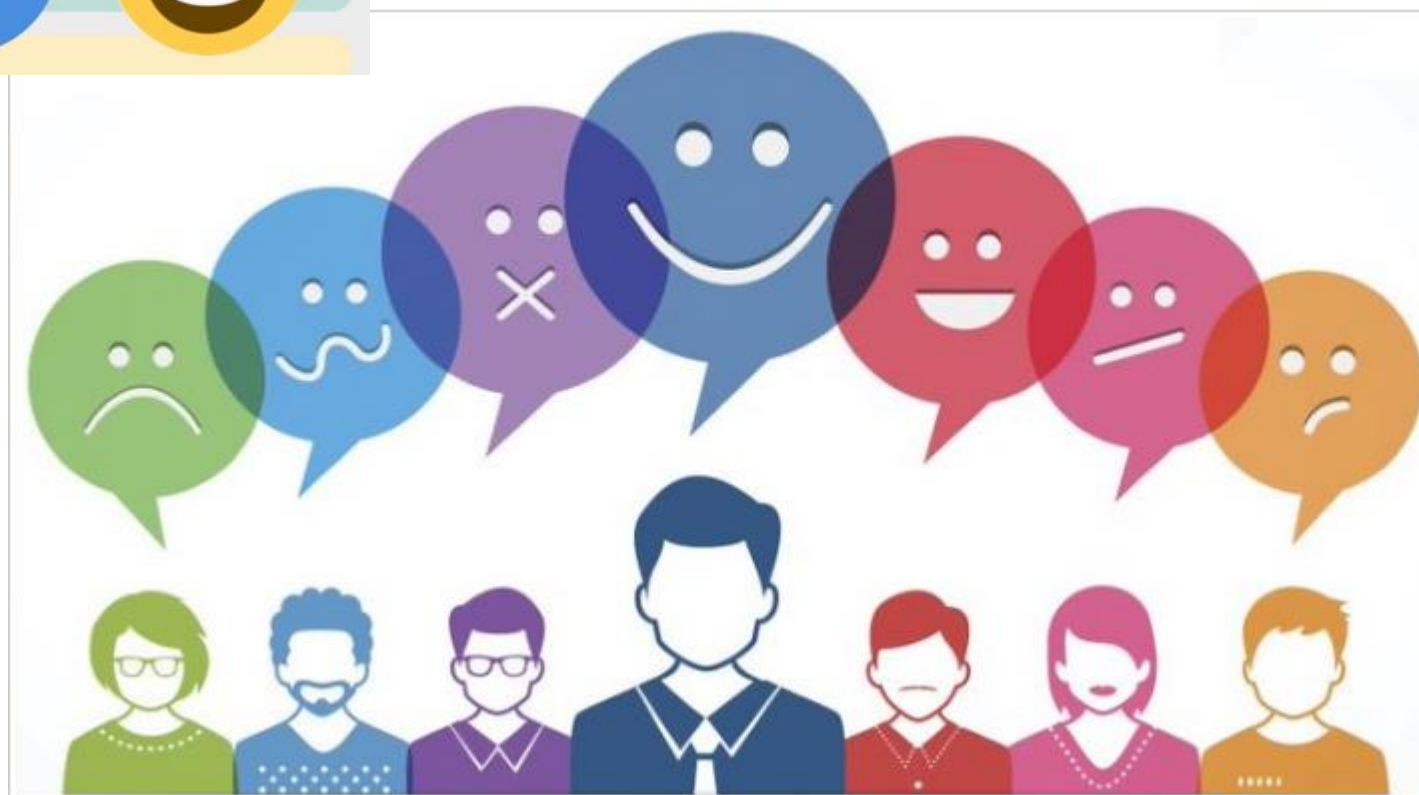
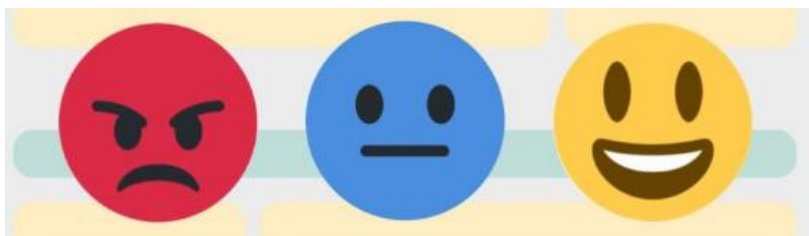
Where: Gates 159

Hamidreza Baradaran Kashani



کاربردهای NLP

❖ تحلیل احساسات (Sentiment Analysis)

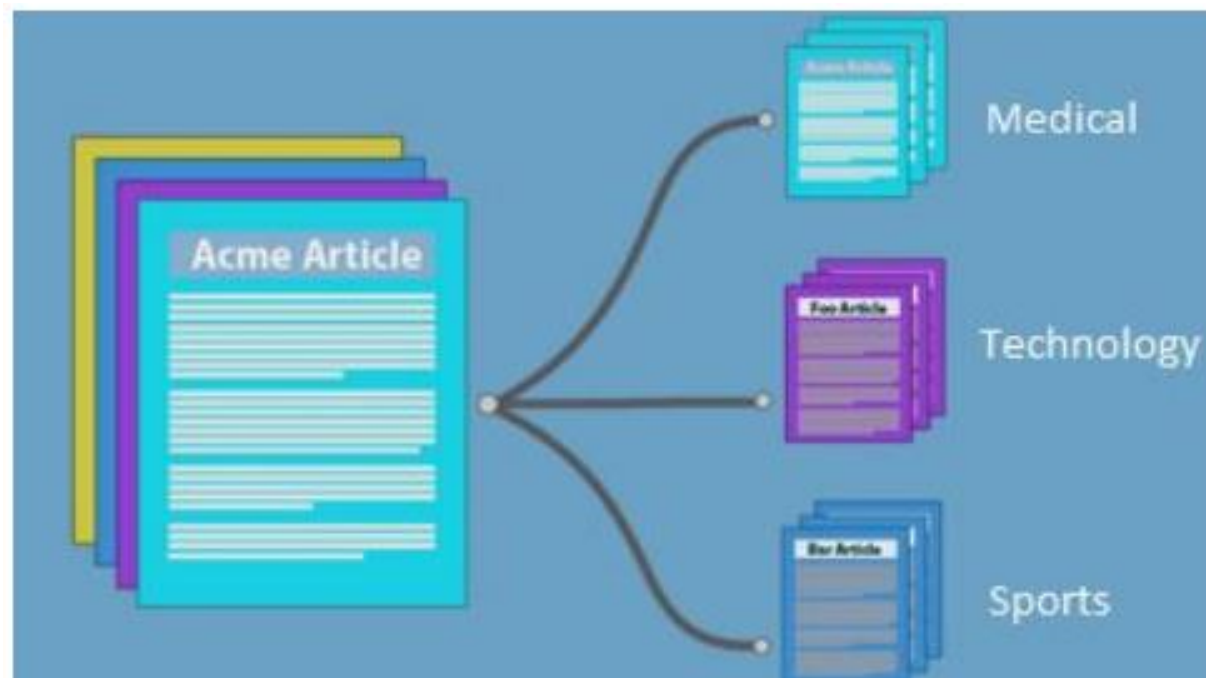


Hamidreza Baradaran Kashani



کاربردهای NLP

❖ دسته بندی متون (Text Classification)



Hamidreza Baradaran Kashani



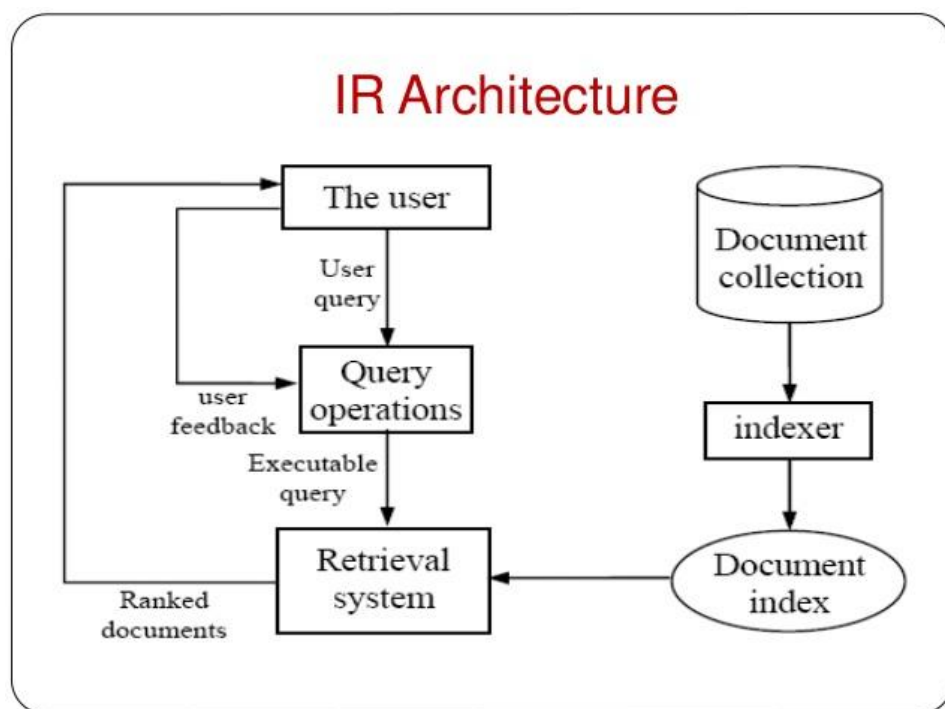
کاربردهای NLP

❖ بازیابی اطلاعات (Information Retrieval)

❖ بازیابی اطلاعات ممکن است به عنوان یک برنامه نرم افزاری تعریف شود که به سازماندهی، ذخیره سازی، بازیابی و ارزیابی اطلاعات از مخازن اسناد به ویژه اطلاعات متنی می پردازد.

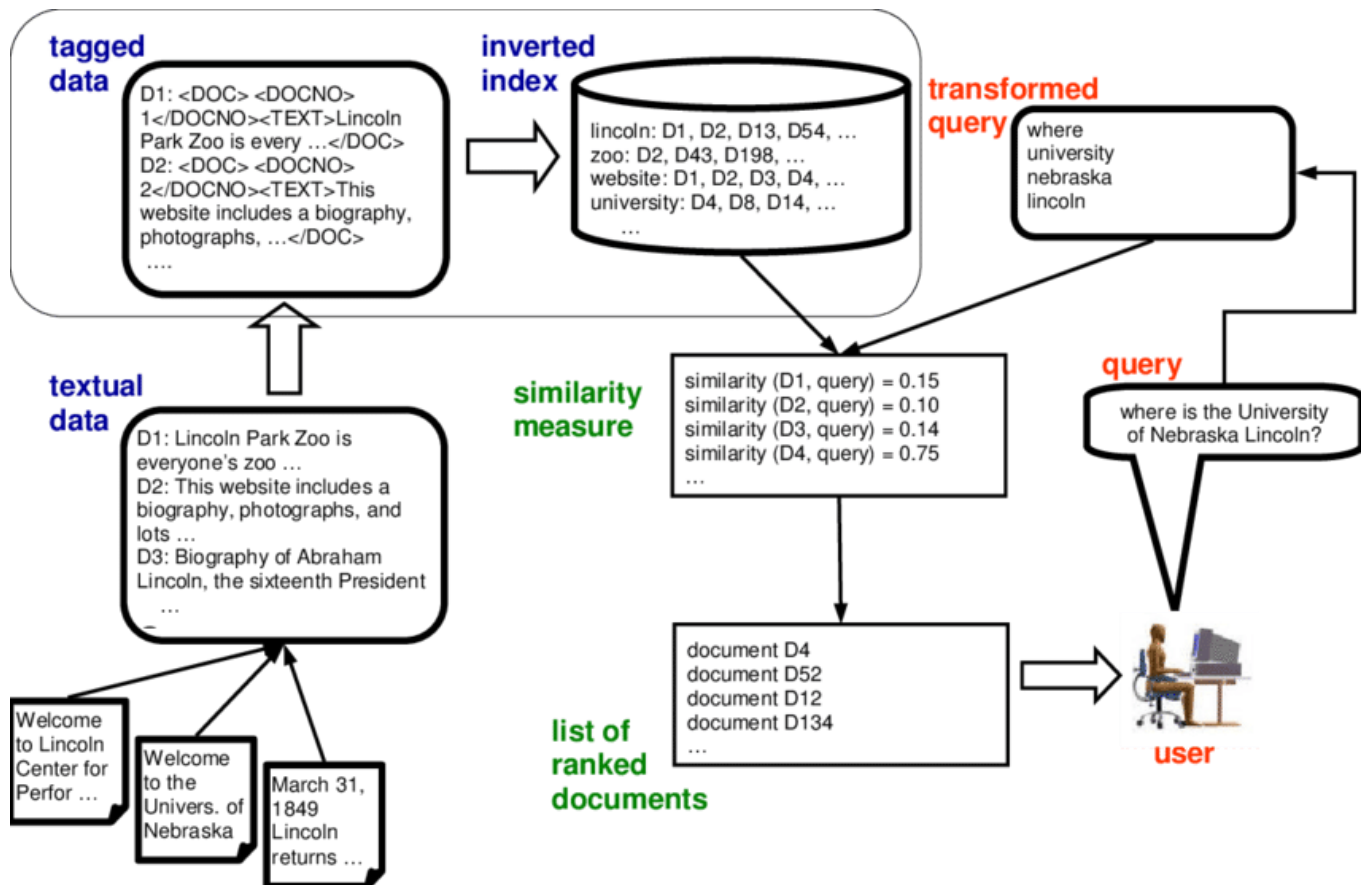
❖ این سیستم به کاربران در یافتن اطلاعات مورد نیاز کمک می کند اما به صراحت جواب سوالات را بر نمی گرداند.

❖ اسنادی که نیاز کاربر را برآورده کند، اسناد مرتبط (relevant documents) نامیده می شوند.





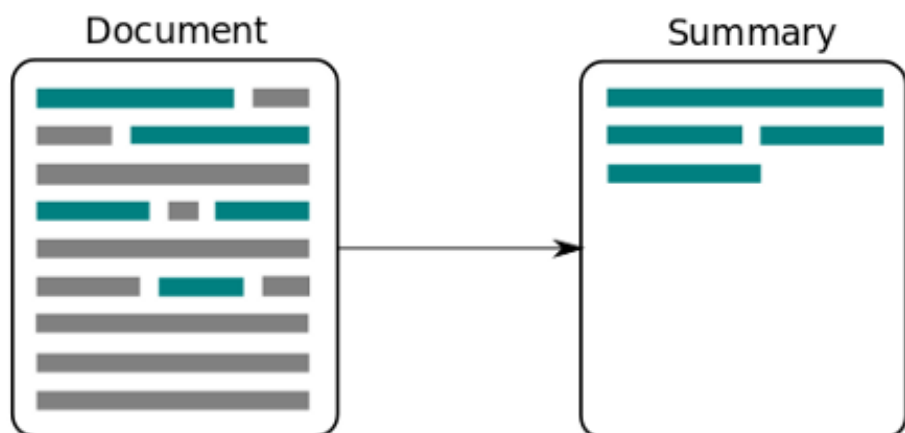
❖ بازیابی اطلاعات (Information Retrieval) – ادامه



Hamidreza Baradaran Kashani



❖ خلاصه سازی متون (Text Summarization)



GOLD

Gas mileage is below what is expected. Gas mileage is below what is expected.
Gas mileage is good once car is broken in.
The gas mileage on the camry was not good as it was advertised to be.

This car looks great and the build quality is good. Mine suffers from the tranny slip on the 3, 4 upshift when cold, build quality is good but nothing like the Toyota, hype I was expecting, as it has its share of squeaks and rattles just like a 10 year old Chevy. Overall a good car, no build quality issues yet. This was the 7th Camry and the last, I feel Toyota quality is slipping big time.

EXTRACTIVE

ABSTRACTIVE

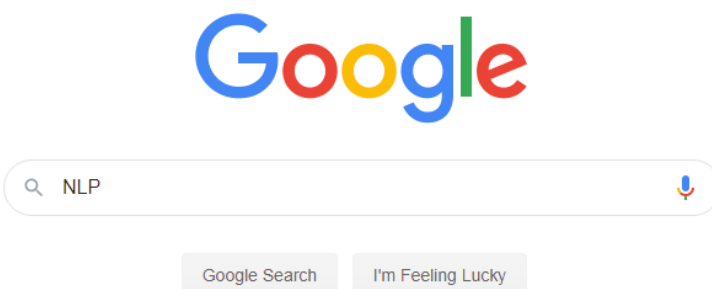
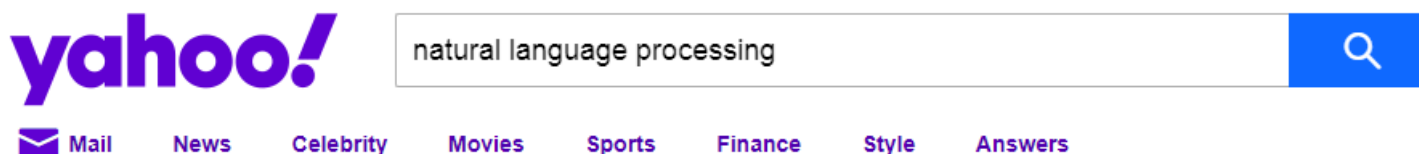
The gas mileage is still good. Cant give specifics but for v6, its good. Being a mother who drives a lot i wanted a safe vehicle with good gas mileage. The interior is roomy, the ride is smooth and solid and yet it has excellent gas mileage. The corolla is the best toyota car when it comes to gas per mileage



کاربردهای NLP

❖ موتورهای جستجو (Search Engines)

- ❖ مانند Google، Yahoo و Bing یا موتورهای جستجو مشهور در زبانهای دیگر مثل Baidu در چینی
- ❖ تمام این موتورها از روش های NLP بهره می برند بطوریکه سعی کرده تا مورد های جستجو و سوالات کاربر (queries) را درک کرده (understanding) و سندهای تطبیق یافته با آن مورد را برای کاربر بیابند.



Hamidreza Baradaran Kashani



کاربردهای NLP

❖ پاسخ سوالات (Question Answering: QA)

❖ برنده یک مسابقه تلویزیونی در ۱۶ فوریه ۲۰۱۱: سیستم مشهور IBM's watson



Hamidreza Baradaran Kashani



❖ پاسخ سوالات مبتنی بر وب (Web-based Question Answering)

❖ یک تعمیم از جستجوی وب ساده است که در آن کاربر به جای آن که فقط کلمات کلیدی (keywords) را تایپ کند، می تواند سوالات کاملی را از سوالات ساده تا سخت مطرح کند، مثلاً؟

“what does “divergent” mean?”

“What year was Abraham Lincoln born?”

How much Chinese silk was exported to Western Europe by the end of the 18th century?

«دانشگاه های مختلف ایران در چه حوزه هایی از پردازش زبان طبیعی تحقیقات بیشتری داشته اند؟»

❖ برای سوالاتی که حالت تعریف (definition questions) داشته یا تنها برخی فاکتورها مثل زمان یا مکان یک رخداد را اشاره دارد، معمولاً همان موتور جستجو پاسخ کاربر را می دهد. اما برای سوالات پیچیده که معمولاً نیاز به یک سری استنتاج ها بر اساس تعدادی واقعیت (facts) دارد، قطعاً روش های پردازش زبان طبیعی پیچیده تری نیاز است.

Hamidreza Baradaran Kashani



کاربردهای NLP

❖ ترجمه ماشینی (Machine Translation)

Google Translate

Google Translate interface showing a translation from English to Persian.

Text: Machine translation is another important new application and machine translation can be fully automatic. So for example, we might have a source sentence in English or Chinese and here's Stanford's phrasal MT system translating that into Persian. But MT can also be used to help human translators.

Persian Translation: ترجمه ماشینی یکی دیگر از مهمترین کاربردهای جدید است و ترجمه ماشینی می تواند کاملاً اتوماتیک باشد. برای مثال ، ممکن است یک جمله مبدا به انگلیسی یا چینی داشته باشیم و در اینجا سیستم MT MT عبارت استوفورد که آن را به فارسی ترجمه می کند. اما MT نیز می تواند برای کمک به مترجمان انسانی استفاده شود.

Send feedback

Hamidreza Baradaran Kashani



کاربردهای NLP

❖ سیستم های گفتگو (Dialog Systems)

User: I need a flight from Tehran to Isfahan, arriving by 9 pm?

System: What day are you flying on?

User: Tomorrow

System: Returns a list of flights

Hamidreza Baradaran Kashani



making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation

I need new batteries for my **mouse**.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party May 27 add

mostly solved

Spam detection

Let's go to Agra!

Buy V1AGRA ...

Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Hamidreza Baradaran Kashani





چرا NLP سخت است؟

- ❖ عدم رعایت قواعد گرامری توسط افراد مختلف یک زبان،
- ❖ انعطاف پذیری گرامری بالا در زبان های مختلف،
- ❖ عدم استفاده افراد از علائم نقطه گذاری در جملات،
- ❖ مثلاً جمله زیر را در نظر بگیرید که معانی مختلفی از آن برداشت می شود (گرامر آن کاملاً صحیح است):

“At last a computer that understands you like your mother”

1. *It understands you as well as your mother understands you*
2. *It understands that you like your mother*
3. *It understands you as well as it understands your mother*

Hamidreza Baradaran Kashani



چرا NLP سخت است؟

❖ ابهام (ambiguity) در سطوح مختلفی ممکن است حاصل شود:

❖ ابهام در سطح آکوستیکی (acoustic level)

○ مثلاً یک سیستم بازشناسی گفتار (Speech Recognizer) ممکن است با شنیدن جمله فوق دو نتیجه زیر را گزارش دهد:

1. ... a computer that understand you **like your** mother
2. ... a computer that understand you **lie cured** mother

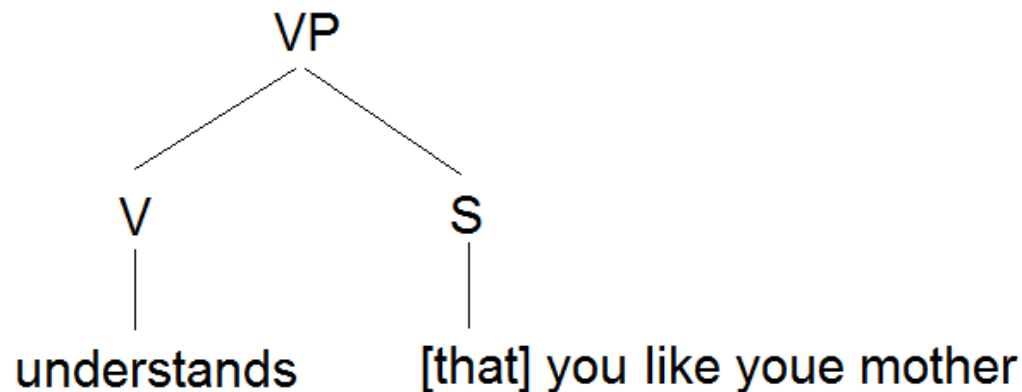
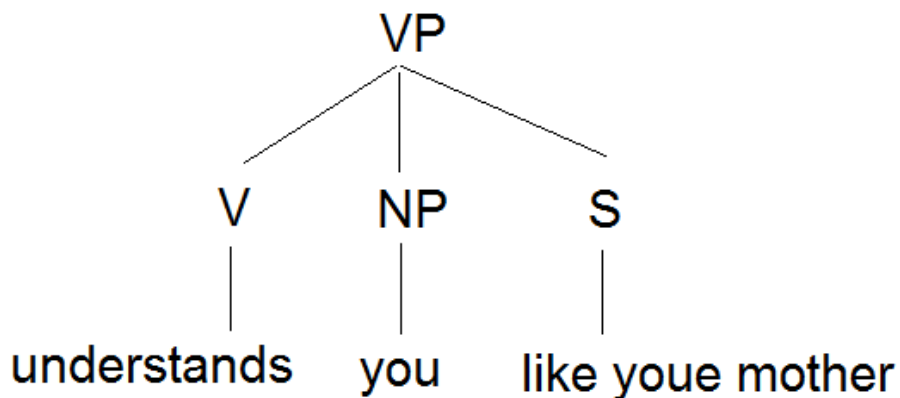
❖ هر دو عبارت **like your** و **lie cured**، دو دنباله دو کلمه ای معتبر از لحاظ گرامری در زبان انگلیسی هستند، با این حال بواسطه شباهت تلفظ آکوستیکی دو دنباله ممکن است سیستم بازشناسی در تشخیص متن صحیح دچار مشکل شود، اگرچه با استفاده از مدل‌های زبانی (علاوه بر مدل آکوستیکی) احتمال بالاتری به عبارت اول داده می‌شود.



چرا NLP سخت است؟

❖ ابهام در سطح نحوی (syntactic level)

- ممکن است یک الگوریتم تجزیه گر نحوی (syntactic parser) دو تجزیه مختلف را برای یک جمله حاصل کند.



- ساختارهای نحوی مختلف منجر به تعبیرهای مختلفی هم خواهد شد.



چرا NLP سخت است؟

❖ ابهام در سطح معنایی (semantic level)

○ این ابهام را می توان نمونه ای از ابهام معنی (حس) کلمه (word sense ambiguity) در نظر گرفت.

○ در واقع در دیکشنری هر زبان کلمات زیادی وجود دارند که برای آنها چند معنی بیان شده است، بسته به آنکه آن کلمه در چه متنی یا دنباله ای از کلمات مجاورش (context) بکار رفته است.

○ برای مثال ابهام در معنی کلمه شیر در جمله:

○ من یک شیر خریدم.





چرا NLP سخت است؟

❖ ابهام در سطح گفتمان یا چند بند (discourse level or multi-clause)

- یک گفتمان در حالت کلی از چند جمله می تواند تشکیل شود.
- در جملات مختلف و پیوسته معمولاً ضمایی استفاده می شود که به برخی از موجودیت ها در همان جمله و یا جملات قبل مرتبط می شود. این مساله، یافتن مرجع ضمائر یا co-reference resolution نامیده می شود.

○ برای مثال:

علی به مغازه پدرش رفت. او معمولاً دوشنبه ها به آنجا می رود.



مهارت های مورد نیاز

❖ آشنایی با موارد زیر:

- جبر خطی (بردارها، ماتریس ها و عملیات روی آن ها)
- تئوری احتمالات پایه
- برنامه نویسی بویژه پایتون
- آشنایی با مفاهیم یادگیری ماشین

Hamidreza Baradaran Kashani