بازیابی اطلاعات

دکتر امین گلزاری اسکوئی

a.golzari@azaruniv.ac.ir

a.golzari@tabrizu.ac.ir

https://github.com/Amin-Golzari-Oskouei

دانشگاه صنعتی ارومیه

پاییز ۱۴۰۲

فصل ۶

امتیازدهی، وزن دهی عبارات و مدل فضای بردار

مطالب این فصل

- رتبه بندی نتایج جستجو
- Term frequency
- Tf-idf ranking
- Vector space model

بازیابی رتبه بندی شده

- تا اینجای کار، پرس‌وجوهای ما همگی بصورت بولی بوده اند.
  - اسناد یا مطابقت دارند یا ندارند.
- این برای کاربران حرفه ای که درک دقیقی از نیازهای خود و مجموعه دارند خوب است.
- برای برنامه ها نیز خوب است: برنامه ها به راحتی می توانند هزارن نتیجه را مصرف کنند.
- برای اکثر کاربران خوب نیست.
- اکثر کاربران قادر به نوشتن پرس‌وجوهای بولی نیستند. . .
  - یا قادر به نوشتن پرس‌وجوهای بولی هستند، اما فکر می کنند این کار زیادی است.
- بیشتر کاربران نمی‌خواهند از هزاران نتیجه عبور کنند.
- این به ویژه در مورد جستجو وب صادق است.

مشکل جستجوی بولی: یا جشن یا قحطی (یا فراوانی یا هیچ)

- پرس‌وجوهای بولی اغلب به نتایج بسیار کم (=0) یا بیش از حد (هزاران) منجر می‌شوند.
- Query 1 (boolean conjunction): [standard user dlink 650]
- → 200,000 hits – feast
- Query 2 (boolean conjunction): [standard user dlink 650 no card found]
- → 0 hits – famine
- در بازیابی بولی، رسیدن به یک پرس‌وجو که تعداد نتیجه موفق قابل مدیریتی را ایجاد می‌کند، مهارت زیادی می خواهد.

جشن یا قحطی: مشکلی در بازیابی رتبه بندی وجود ندارد

- با رتبه‌بندی، مجموعه‌های نتایج بزرگ مسئله‌ای نیستند.
- صرفا 10 نتیجه برتر را نشان می‌دهد.
- کاربر را غرق در جستجو نمی‌کند.
- فرض: الگوریتم رتبه‌بندی کار می‌کند، نتایج مرتبط تر نسبت به نتایج کمتر مرتبط، بالاتر رتبه‌بندی می‌شوند.

امتیازدهی به عنوان مبنای بازیابی رتبه‌بندی‌شده

- ما می‌خواهیم اسنادی را که مرتبط‌تر هستند، بالاتر از اسنادی که کمتر مرتبط هستند، رتبه‌بندی کنیم.
- چگونه می‌توانیم چنین رتبه بندی اسناد مجموعه را با یک پرس‌وجو مربوطه انجام دهیم؟
- به هر جفت(پرس‌وجو و سند) یک امتیاز اختصاص دهید، مثلاً بین 1 و 0
- این امتیاز میزان طتابق سند و پرس‌وجو را اندازه گیری می کند.

امتیازهای تطبیق پرس‌وجو–سند

- چگونه امتیاز یک جفت پرس‌وجو-سند را حساب کنیم؟
- بیایید با یک پرس‌وجو تک واژه‌ای شروع کنیم.
- در صورتی که عبارت پرس‌وجو در سند وجود نداشته باش؛ امتیاز باید 0 باشد.
- هر چه عبارت پرس‌وجو در سند بیشتر باشد، امتیاز بالاتری خواهد داشت.
- ما به تعدادی از روش‌های جایگزین برای انجام این کار نگاه خواهیم کرد.

روش اول: ضریب جاکارد

- یک معیار رایج برای سنجش میزان همپوشانی دو مجموعه
- فرض کنید A و B دو مجموعه هستند
- ضریب جاکارد:

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$
$$(A \neq \emptyset \text{ or } B \neq \emptyset)$$

- JACCARD $(A, A) = 1$
- JACCARD $(A, B) = 0$ if $A \cap B = 0$
- A و B نیاز نیست هم اندازه باشند.
- همیشه عددی بین 0 و 1 اختصاص می دهد.

ضریب جاکارد: مثال

❖ امتیاز تطابق پرس‌وجو و سند که ضریب جاکارد برای آن محاسبه می‌کند، چقدر است؟

- Query: “dies of March”
- Document “Caesar died in March”
- JACCARD(q, d) = 1/6

ایراد جاکارد چیست؟

- میزان تکرار عبارت را در نظر نمی‌گیرد(یک اصطلاح چند تکرار دارد؟)
- اصطلاحات نادر آموزنده‌تر از اصطلاحات متداول هستند.
- جاکارد این بحث را در نظر نمی‌گیرد.
- ما به روش پیچیده‌تری برای عادی سازی طول یک سند نیاز داریم.
- بعداً در این بحث، به جای جاکارد برای عادی سازی طول از $|A \cap B| / \sqrt{|A \cup B|}$ (کسینوس) استفاده خواهیم کرد، به جای $|A \cap B| / |A \cup B|$ به جای عادی سازی طول.

مدل استقلال دودویی

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth . . .
ANTHONY	1	1	0	0	0	1
BRUTUS	1	1	0	1	0	0
CAESAR	1	1	0	1	1	1
CALPURNIA	0	1	0	0	0	0
CLEOPATRA	1	0	0	0	0	0
MERCY	1	0	1	1	1	1
WORSER	1	0	1	1	1	0
. . .						

- هر سند به صورت یک بردار باینری $\{0, 1\}^{|V|}$ نشان داده می‌شود.

مدل استقلال دودویی

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth . . .
ANTHONY	157	73	0	0	0	1
BRUTUS	4	157	0	2	0	0
CAESAR	232	227	0	2	1	0
CALPURNIA	0	10	0	0	0	0
CLEOPATRA	57	0	0	0	0	0
MERCY	2	0	3	8	5	8
WORSER	2	0	1	1	1	5
. . .						

- هر سند به صورت یک بردار باینری $\mathrm{N}^{|V|}$ نشان داده می‌شود.

مدل کیسه کلمات

- ما در یک سند ترتیب کلمات را در نظر نمی‌گیریم.
- *John is quicker than Mary and Mary is quicker than John*
- به ترتیب یکسان نشان داده می‌شوند.
- به این، مدل کیسه کلمات می‌گویند.
- به تعبیری، این یک گام به عقب است: شاخص موقعیتی توانست این دو سند را از هم متمایز کند.
- ما بعداً در این دوره به "بازیابی" اطلاعات موقعیتی خواهیم پرداخت.
- اما فعلا توجه‌مان روی مدل کیسه کلمات خواهد بود.

Term frequency tf

- تناوب عبارت $tf_{t,d}$ از عبارت t در سند d به عنوان تعداد دفعاتی که t در d رخ می دهد تعریف می‌شود.
- ما می خواهیم از tf هنگام محاسبه امتیازهای تطبیق پرس‌وجو و سند استفاده کنیم.
- اما چگونه؟
- فرکانس عبارت خام آن چیزی نیست که ما می‌خواهیم زیرا:
- سندی با $tf = 10$ وقوع عبارت مرتبط تر از سندی با $tf = 1$ وقوع عبارت است.
- اما 10 برابر مرتبط تر نیست.
- میزان مرتبط بودن با میزان تکرار افزایش نمی‌یابد.

جایگزین تناوب خام: وزن تناوب لگاریتمی

- وزن فرکانس لگاریتم عبارت t در d به صورت زیر تعریف می شود.

$$w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d} & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$\mathrm{tf}_{t,d} \rightarrow w_{t,d}$: $0 \rightarrow 0,\ 1 \rightarrow 1,\ 2 \rightarrow 1.3,\ 10 \rightarrow 2,\ 1000 \rightarrow 4,$ etc.

- امتیاز برای یک جفت سند-پرس‌وجو: جمع بر عبارات t در q و d

- اگر هیچ یک از عبارت های پرس‌وجو در سند وجود نداشته باشد، امتیاز 0 است.

تمرین

- امتیاز تطبیق جاکارد و امتیاز تطبیق tf را برای جفت‌های پرسش و سند زیر محاسبه کنید.

- q: [information on cars] d: “all you’ve ever wanted to know about cars”
- q: [information on cars] d: “information on trucks, information on planes, information on trains”
- q: [red cars and red trucks] d: “cops stop red cars more often”

تناوب در سند در مقابل تناوب در مجموعه

- علاوه بر فراوانی عبارت (تکرار عبارت در سند) . . .
- . . همچنین می خواهیم از فراوانی عبارت در مجموعه برای وزن دهی و رتبه بندی استفاده کنیم.

وزن مطلوب برای عبارات نادر

- عبارات نادر و کمیاب آموزنده تر از عبارات پر تکرار هستند.
- عبارتی را در پرس‌وجو در نظر بگیرید که در مجموعه نادر است (به عنوان مثال، ARACHNOCENTRIC.)
- سند حاوی این اصطلاح به احتمال زیاد مرتبط است.
- ما برای عبارات کمیاب مانند ARACHNOCENTRIC وزن های بالا می خواهیم.

وزن مطلوب برای عبارات نادر

- عبارت متداول نسبت به عبارات نادر اطلاعات کمتری دارند.
- عبارتی را در پرس و جو در نظر بگیرید که در مجموعه مکرر است (به عنوان مثال، GOOD، INCREASE، LINE).سندی که حاوی این اصطلاح است بیشتر از سندی که ندارد مرتبط است. . .
- . . . اما کلماتی مانند GOOD،INCREASE و LINE شاخص های مطمئنی برای میزان ارتباط نیستند.
- ← برای عبارات مکرر مانند GOOD، INCREASE و LINE، ما وزن های مثبت می‌خواهیم. . .
- . . . اما با وزن کمتری نسبت به عبارات نادر.

تناوب سند

- ما برای اصطلاحات کمیاب مانند ARACHNOCENTRIC وزن های بالا می‌خواهیم.
- ما وزن کم(مثبت)را برای کلمات متداول مانند GOOD، INCREASE و LINE می‌خواهیم.
- ما از تناوب سند استفاده خواهیم کرد تا این را در محاسبه امتیاز تطبیق لحاظ کنیم.
- فراوانی سند، تعداد اسناد موجود در مجموعه ای است که این عبارت در آن وجود دارد.

وزن idf

- dft تناوب سند است، تعداد اسنادی که t در آنها رخ می‌دهد.
- dft معیار معکوس میزان اطلاعات مفید عبارت t است.
- وزن idf عبارت t را به صورت زیر تعریف می‌کنیم:

$$\text{idf}_t = \log_{10} \frac{N}{\text{df}_t}$$

(N) تعداد اسناد موجود در مجموعه است.

- idft معیاری برای آموزنده بودن این عبارت است.
- [$\log N/\text{df}_t$] به جای [N/df_t] برای "کم کردن" اثر idf
- توجه داشته باشید که ما از تبدیل لگاریتمی برای تناوب عبارت و تناوب سند استفاده می کنیم.

مثالی برای idf

❖ با استفاده از فرمول روبه‌رو داریم :

$$\text{idf}_t = \log_{10} \frac{1{,}000{,}000}{\text{df}_t}$$

term	df_t	idf_t
calpurnia	1	6
animal	100	4
sunday	1000	3
fly	10,000	2
under	100,000	1
the	1,000,000	0

تاثیر idf بر رتبه بندی

- idf بر رتبه بندی اسناد برای پرس‌وجوها با حداقل دو عبارت تأثیر می‌گذارد.
- به عنوان مثال، در پرس‌وجو "arachnocentric line"، وزن دهی idf وزن نسبی ARACHNOCENTRIC را افزایش می‌دهد و وزن نسبی LINE را کاهش می‌دهد.
- idf تأثیر کمی بر رتبه بندی برای پرس‌وجوهای تک عبارتی دارد.

تناوب در سند در مقابل تناوب در مجموعه

کلمه	تناوب در مجموعه	تناوب در سند
INSURANC	10440	3997
E	10422	8760
TRY		

- تناوب مجموعه t: تعداد توکن های t موجود در مجموعه
- فراوانی سند t: تعداد اسنادی که t در آن وجود دارد.
- چرا این ارقام؟
- کدام کلمه عبارت جستجوی بهتری است (و باید وزن بیشتری داشته باشد)؟
- این مثال نشان می دهد که تناوب در سند برای وزن دهی بهتر از تناوب در مجموعه است.

وزن دهی tf-idf

- وزن tf-idf یک عبارت حاصل ضرب وزن tf و وزن idf آن است.

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \cdot \log \frac{N}{\mathrm{df}_t}$$

- tf-weight
- idf-weight

- شناخته شده ترین طرح وزن دهی در بازیابی اطلاعات
- توجه: "-" در tf-idf یک خط فاصله است، نه علامت منفی!
- نام های دیگر: tf.idf، tf x idf

خلاصه: tf-idf

- یک وزن tf-idf برای هر عبارت t در هر سند d تعیین میکنیم:

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \cdot \log \frac{N}{\mathrm{df}_t}$$

- وزن tf-idf با تعداد رخدادها در یک سند (فراوانی عبارت)

- ... و همچنین با نادر بودن این واژه در مجموعه افزایش می یابد. (فراوانی سند وارونه)

تمرین: تکرر عبارت، مجموعه و سند

کمیت	نماد	تعریف
فراوانی عبارت	$tf_{t,d}$	تعداد رخداد t در d
فراوانی سند	df_t	تعداد اسناد موجود در مجموعه ای که t در آنها وجود دارد
فراوانی مجموعه	cf_t	تعداد کل رخداد t در مجموعه

- رابطه بین df و cf؟
- رابطه tf و cf؟
- رابطه tf و df؟

مدل استقلال دودویی

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth . . .
ANTHONY	1	1	0	0	0	1
BRUTUS	1	1	0	1	0	0
CAESAR	1	1	0	1	1	1
CALPURNIA	0	1	0	0	0	0
CLEOPATRA	1	0	0	0	0	0
MERCY	1	0	1	1	1	1
WORSER	1	0	1	1	1	0
. . .						

- هر سند به صورت یک بردار باینری $\{0, 1\}^{|V|}$ نشان داده می‌شود.

Count matrix

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth . . .
ANTHONY	157	73	0	0	0	1
BRUTUS	4	157	0	2	0	0
CAESAR	232	227	0	2	1	0
CALPURNIA	0	10	0	0	0	0
CLEOPATRA	57	0	0	0	0	0
MERCY	2	0	3	8	5	8
WORSER	2	0	1	1	1	5
. . .						

- هر سند به صورت یک بردار باینری $\mathrm{N}^{|V|}$ نشان داده می‌شود.

Binary → count → weight matrix

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth . . .
ANTHONY	5.25	3.18	0.0	0.0	0.0	0.35
BRUTUS	1.21	6.10	0.0	1.0	0.0	0.0
CAESAR	8.59	2.54	0.0	1.51	0.25	0.0
CALPURNIA	0.0	1.54	0.0	0.0	0.0	0.0
CLEOPATRA	2.85	0.0	0.0	0.0	0.0	0.0
MERCY	1.51	0.0	1.90	0.12	5.25	0.88
WORSER	1.37	0.0	0.11	4.15	0.25	1.95
. . .						

هر سند اکنون به عنوان یک بردار با ارزش واقعی از $\text{tf idf weights} \in R^{|V|}$ نشان داده می شود.

اسناد به عنوان بردار

- اکنون هر سند به عنوان یک بردار با ارزش واقعی وزن‌های tf-idf نشان داده می‌شود $\in R^{|V|}$
- بنابراین یک فضای برداری با ارزش واقعی $|V|$ابعدی داریم.
- عبارات محورهای فضا هستند.
- اسناد نقاط یا بردارهایی در این فضا هستند.
- ابعاد بسیار بالا: وقتی این را در موتورهای جستجوی وب اعمال می‌کنید
- ده‌ها میلیون بعد ایجاد می‌شود.
- هر بردار بسیار پراکنده است - بیشتر ورودی ها صفر هستند.

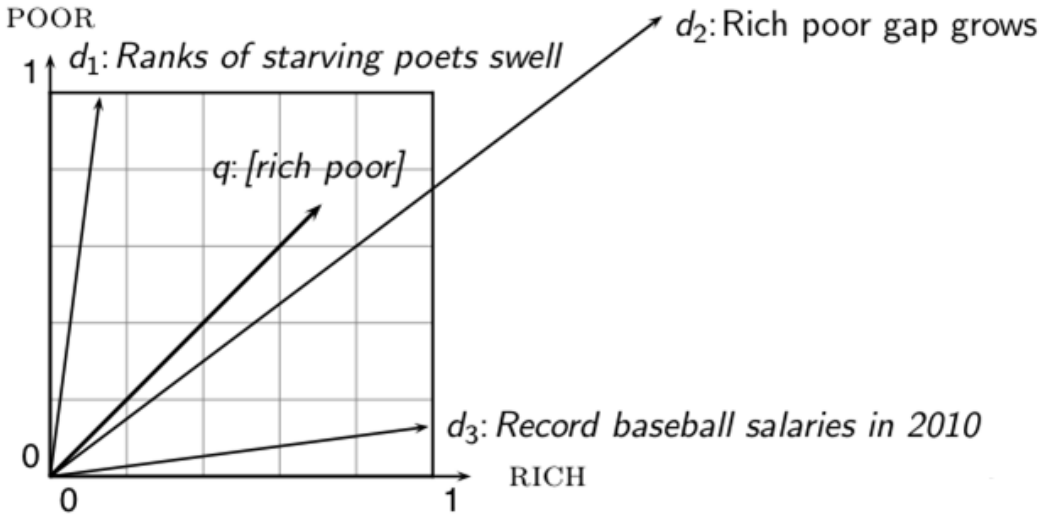
پرس‌وجو به‌عنوان بردار

- ایده کلیدی 1: همین کار را برای پرس‌وجوها انجام دهید: آنها را به عنوان بردار در فضای با ابعاد بالا نشان دهید.
- ایده کلیدی 2: اسناد را بر اساس نزدیکی و مطابقت آنها به پرس‌وجو رتبه بندی کنید.
- میزان دقیق بودن = شباهت
- مجاورت ≈ فاصله منفی
- به یاد داشته باشید ما این کار را انجام می‌دهیم چون می‌خواهیم مدل بولی همه یا هیچ، جشن یا قحطی را کنار بگذاریم.
- به جای آن، اسناد مربوطه را بالاتر از اسناد غیر مرتبط قرار می‌دهیم.

چگونه شباهت فضای برداری را رسمی یا فرمالیته کنیم؟

- برش اول: (منفی)فاصله بین دو نقطه
- (= فاصله بین نقاط انتهایی دو بردار)
- فاصله اقلیدسی؟
- فاصله اقلیدسی ایده بدی است چون فاصله اقلیدسی برای بردارهایی با طول های مختلف زیاد است.

چرا فاصله فکر بدی است؟



فاصله اقلیدسی $\vec{q}$ و $\vec{d_2}$ بزرگ است اگرچه توزیع عبارات در پرس‌وجو q و توزیع اصطلاحات در سند d2 بسیار شبیه هستند.

سؤالاتی در مورد تنظیم فضای برداری پایه؟

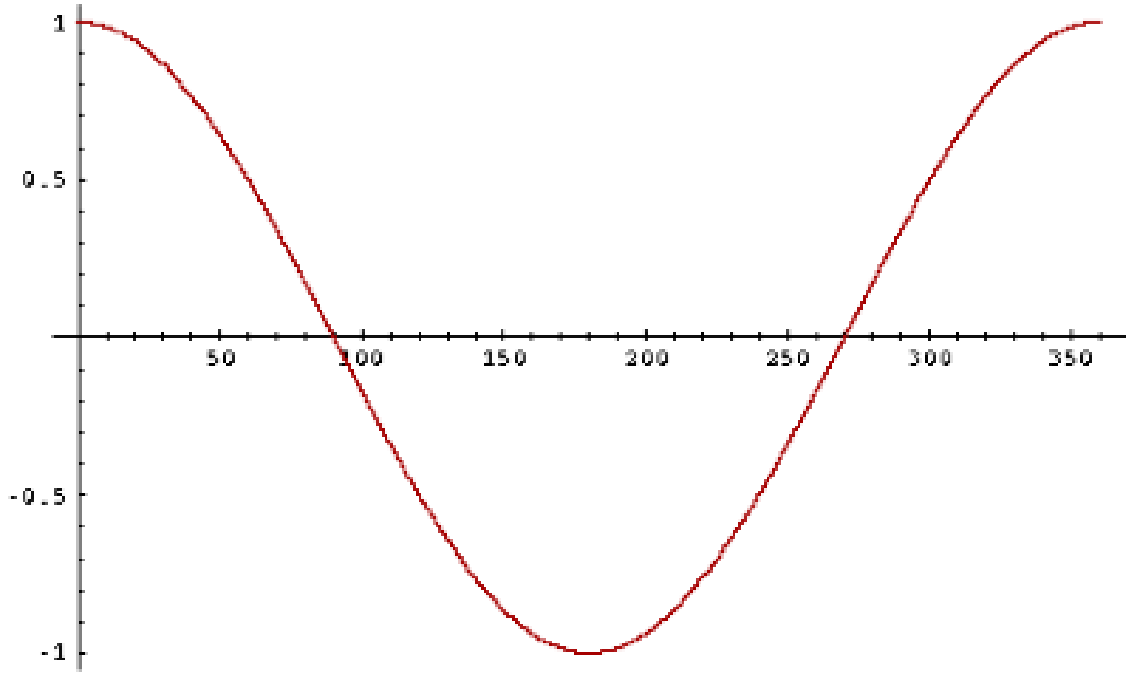
به جای فاصله از زاویه استفاده کنید

- اسناد را بر اساس زاویه به وسیله پرس‌وجو رتبه بندی کنید
- آزمایش فکری: یک سند d را بردارید و به خودش اضافه کنید. این سند را d′ بنامید. d’ دو برابر d است.
- "از نظر معنایی" d و d’ دارای محتوای یکسانی هستند.
- زاویه بین دو سند 0 است که مربوط به حداکثر شباهت است حتی اگر فاصله اقلیدسی بین دو سند بسیار زیاد باشد.

از زاویه تا کسینوس

- دو مفهوم زیر معادل هستند.
- اسناد را با توجه به زاویه بین پرس‌وجو و سند به ترتیب کاهشی رتبه بندی کنید.
- اسناد را بر اساس کسینوس (پرس و جو، سند) به ترتیب افزایش رتبه بندی کنید.
- کسینوس یک تابع کاهنده یکنواخت زاویه برای بازه [0°, 180°] است.

کسینو



نرمال سازی طول

- چگونه کسینوس را محاسبه کنیم؟
- یک بردار را می توان با تقسیم هر یک از اجزای آن بر طول آن نرمال کرد - در اینجا از نرم L2 استفاده می‌کنیم:

$$||x||_2 = \sqrt{\sum_i x_i^2}$$

- این بردارها را پس از نرمال‌سازی بر روی کره واحد نشان می‌دهد:

$$||x||_2 = \sqrt{\sum_i x_i^2} = 1.0$$

- در نتیجه، اسناد طولانی‌تر و اسناد کوتاه‌تر دارای وزن‌هایی به همان ترتیب اندازه هستند.
- تأثیر بر دو سند d و d’ (d پیوست شده به خود) از اسلاید قبلی: آن‌ها پس از نرمال سازی طول بردارهای یکسانی دارند.

شباهت کسینوس بین پرس و جو و سند

$$\cos(\vec{q},\vec{d}) = \text{SIM}(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- qiوزن tf-idfعبارت i در پرس‌وجو است.
- diوزن tf-idfعبارت i در سند است.
- | | و | | طول های و هستند.
- این شباهت کسینوسی و یا به طور معادل، کسینوس زاویه بین و است.

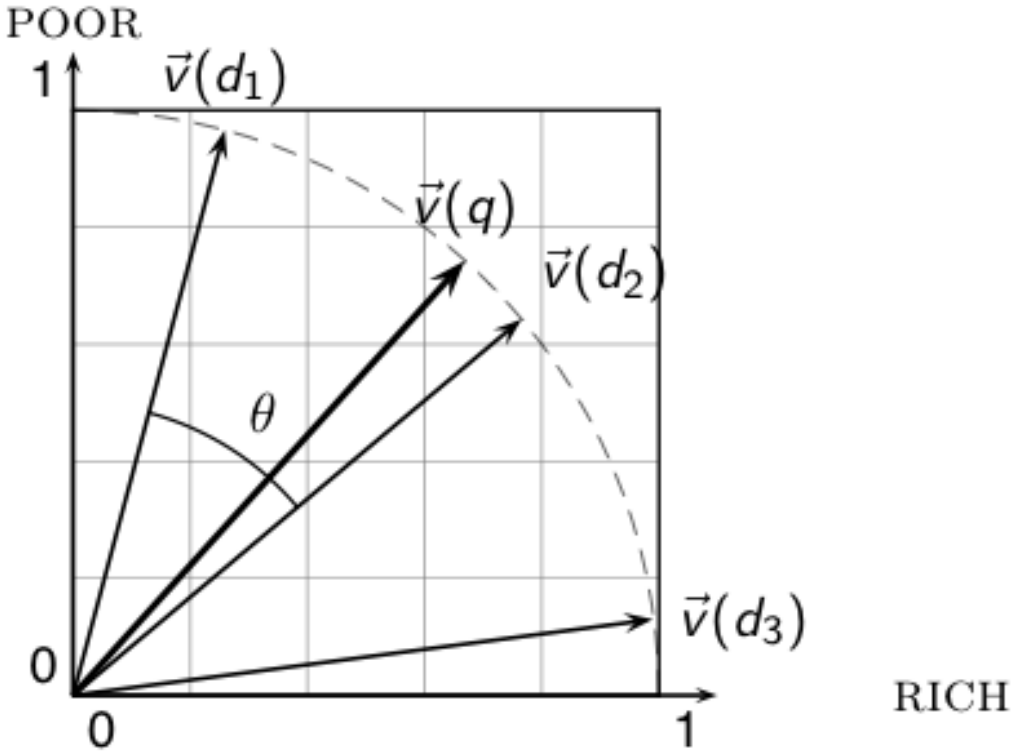
کسینوس برای بردارهای نرمال شده

- برای بردارهای نرمال شده، کسینوس معادل حاصل ضرب نقطه ای یا اسکالری است.

$$\cos(\vec{q}, \vec{d}) = \vec{q} \cdot \vec{d} = \sum_i q_i \cdot d_i$$

- (اگر $\vec{q}$ و $\vec{d}$ به صورت طولی نرمال شده باشند.)

شباهت کسینوسی



مثال :

How similar are these novels? SaS: Sense and Sensibility PaP: Pride and Prejudice WH: Wuthering Heights

term frequencies (counts)

term	SaS	PaP	WH
AFFECTION	115	58	20
JEALOUS	10	7	11
GOSSIP	2	0	6
WUTHERING	0	0	38

مثال :

term frequencies (counts)

term	SaS	PaP	WH
AFFECTION	115	58	20
JEALOUS	10	7	11
GOSSIP	2	0	6
WUTHERING	0	0	38

log frequency weighting

term	SaS	PaP	WH
AFFECTION	3.06	2.76	2.30
JEALOUS	2.0	1.85	2.04
GOSSIP	1.30	0	1.78
WUTHERING	0	0	2.58

برای ساده کردن مثال idf weighting را انجام نمی‌دهیم.

محاسبه امتیاز کسینوس

```
CosineScore(q)
 1  float Scores[N] = 0
 2  float Length[N]
 3  for each query term t
 4  do calculate w_{t,q} and fetch postings list for t
 5      for each pair(d, tf_{t,d}) in postings list
 6      do Scores[d] += w_{t,d} × w_{t,q}
 7  Read the array Length
 8  for each d
 9  do Scores[d] = Scores[d]/Length[d]
10  return Top K components of Scores[]
```

tf-idf weighting اجزای

Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1+\log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5+\frac{0.5\times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases}1 & \text{if } \mathrm{tf}_{t,d}>0\\ 0 & \text{otherwise}\end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha<1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$				

مثال tf-idf

- ما معمولا از وزن‌های مختلف برای پرسش‌ها و اسناد استفاده می‌کنیم.
- علامت گذاری: ddd.qqq
- مثال: lnc.ltn.
- سند: tf لگاریتمی، بدون وزن دهی df، عادی سازی کسینوس.
- پرس و جو: tf لگاریتمی، idf، بدون نرمال سازی.
- آیا بد نیست سند را با idf وزن نکنید؟

- Example query: “best car insurance”
- Example document: “car insurance auto insurance”

مثال Inc.Itn :tf-idf

Query: “best car insurance”. Document: “car insurance auto insurance”.

word	query					document				product
	tf-raw	tf-wght	df	idf	weight	tf-raw	tf-wght	weight	n’lized	
auto	0	0	5000	2.3	0	1	1	1	0.52	0
best	1	1	50000	1.3	1.3	0	0	0	0	0
car	1	1	10000	2.0	2.0	1	1	1	0.52	1.04
insurance	1	1	1000	3.0	3.0	2	1.3	1.3	0.68	2.04

کلید ستون ها:

tf-raw = تناوب عبارت خام (بدون وزن)،

tf-wght = تناوب عبارت وزن دار لگاریتمی، df = تناوب سند، idf = تناوب معکوس سند،

وزن: وزن نهایی عبارت در پرس و جو یا سند، n’lized= وزن های سند پس از نرمال سازی کسینوس، product - حاصل ضرب وزن درخواست نهایی و وزن سند نهایی

$$\sqrt{1^2 + 0^2 + 1^2 + 1.3^2} \approx 1.92$$

$1/1.92 \approx 0.52$

$1.3/1.92 \approx 0.68$ Final similarity score between query and document: $\sum_i w_{qi} \cdot w_{di} = 0 + 0 + 1.04 + 2.04 = 3.08$

سوال؟

خلاصه: بازیابی رتبه بندی شده در مدل فضای برداری

- پرس‌وجو را به عنوان یک بردار وزنی tf-idf نمایش دهید
- هر سند را به عنوان یک بردار وزنی tf-idf نشان دهید
- شباهت کسینوس بین بردار پرسوجو و هر بردار سند را محاسبه کنید
- اسناد را با توجه به درخواست رتبه بندی کنید
- بالاترین K (به عنوان مثال، K = 10) را به کاربر برگردانید

برداشت این فصل

- رتبه بندی نتایج جستجو: چرا مهم است (در مقابل ارائه مجموعه ای از نتایج بولی نامرتب)
- فراوانی عبارت: این یک عنصر کلیدی برای رتبه بندی است.
- رتبه بندی Tf-idf: شناخته شده ترین طرح رتبه بندی سنتی
- مدل فضای برداری: یکی از مهم ترین مدل های رسمی برای بازیابی اطلاعات (همراه با مدل های بولی و احتمالی)

تشکر

سوال؟

a.golzari@azaruniv.ac.ir

a.golzari@tabrizu.ac.ir

https://github.com/Amin-Golzari-Oskouei