Още малко математика, за да разбираме по-добре статистическите методи за обработка на естествен език.

1. Елементарна теория на вероятностите

Знаем що е *събитие*, наблюдение какъв му е *изходът* (*резултатът*), експеримент или *опит* (например хвърляне на монета). Вече видяхме колко е важно да се подберат добре параметрите за наблюдение. Предполагаме набор от експерименти и резултатите им, в нашия случай наблюдения над маркирани корпуси и крайни дискретни множества от резултати. Дефинирахме P(A), вероятност на събитието A. Да припомним с пример с какви понятия си служихме в лекцията за статистическо отгатване (предсказване) на частите на речта (POS tagging).

Пример 1. Идеална монета се хвърля 3 пъти. Каква е вероятността да се падне 2 пъти ези (E)?
Решение: Пространството на експеримента е от 8 равновероятни състояния:

$$\Omega = \{EEE, EET, ETT, ETE, TEE, TET, TTE, TTT\},$$

тоест вероятността на отделния изход (резултат) е 1/8. Това се нарича нормално разпределение. В 1-вата, 2-рата и 5-тата конфигурация от резултати имаме 2 пъти ези, тоест вероятността за 2 ези е 3 от 8 – 3/8.

В тази лекция ще навлезем малко по-дълбоко в статистическите методи, като илюстрираме използваните понятия и подходи върху задачата за извличане на устойчиви словосъчетания (колокации) от немаркиран текст.

2. Колокации

2.1. Увод и просто филтриране според честотата

Колокации (словосъчетания): Това са изрази, съставени от една или повече думи, които съответстват на неявни конвенции как да се казват нещата. Знаят ги всички носители на езика (научават ги с времето и практиката). Например: казваме *силен чай* (тоест богат на някакви съставки, тук значението на *силен* не е основното *имащ голяма физическа сила*). Още казваме *нежен бриз*, но *лек вятър* или *ветрец* е за предпочитане пред *нежен вятър* въпреки че и *силен бриз*, и *силен вятър* са ОК (докато *слаб вятър* е предимно термин и се употребява в метеорологични прогнози). Като колокации могат да се разглеждат и ред термини, напр. *оръжия за масово унищожение*, никой не би казал *масово-унищожаващи оръжия* или нещо подобно. Наблюдаваме фиксирани предпочитания как се съчетават дадени думи и това явление го има във всеки естествен език. Ако чужденец ви каже *мощен чай* сигурно ще се досетите, че има предвид *силен чай*, но също така ще разберете и че не владее добре разговорния език.

Колокациите се характеризират с ограничена композиционалност (значението им е отчасти композирано от значението на съставящите думи, например при *силен чай* има друг смисъл на *силен*). Идиомите са краен пример на не-композиционалност, например значението на *гушвам букета* не може да се композира от *умирам*. Повечето сложни термини се държат като колокации, с установена синтактична и семантична структура, и се употребяват като здраво споена група от думи, които винаги вървят заедно.

Колокациите са важни в компютърната лингвистика, особено при генерацията и машинния превод за вярно предаване на изказването на чужд език, напр. на английски решенията се правят, а не се вземат (тоест трябва да кажем *to make a decision*, а не *to take a decision*). От човешка гледна точка те са важна част от знанията на чужд език и преподаването му и е хубаво да можем да ги събираме автоматично от текстовете. Забележете, че това не са знания за структурата на езика, а за спецификата на употребата на думите, тъй като думите имат «предпочитана компания» от други думи. Вече сме виждали *конкорданс*, речник с контексти около зададена дума (в лекцията за word-sense disambiguation). Но такъв речник изисква човек да го наблюдава и да прави изводите по него. Тук ще разгледаме техники за автоматично извличане на колокации от текстове. Интуитивната идея е, че ако наблюдаваме достатъчно много текстове, трябва да можем да различим типичните употреби на едни и същи думи заедно.

В лекцията ще се разглеждат примери по корпус от вестника New York Times, от август до ноември 1990. Това са 115МБ текст с около 14 милиона думи. За простота фокусът е върху биграми, калокации от две думи, но може да се търсят по-дълги словосъчетания и такива с променлива дължина.

Честота. Най-простата идея е да броим биграмите от думи и да разгледаме най-честите. Но така попадаме на многобройните срещания на биграми от спомагателни и незначещи думи (не случайно ги наричат стоп-думи!): виж таблица 1. Тук е мястото да споменем така наречения закон на Зипф (Zipf's law): Нека преброим думите в голям корпус и ги наредим по намаляващ ред на честотата на срещането им; тогава за всяка дума $\underline{w}$ с честота $\underline{f}$ имаме пореден номер $\underline{r}$, наречен ранг на думата. Законът на Зипф гласи, че има константа $\underline{k}$ такава, че $\underline{f} * \underline{r} = \underline{k}$. Това например значи, че 50-тата дума в списъка се среща около 3 пъти по-често, отколкото 150-тата дума. Всъщност това не е закон, макар да го наричаме с това название, а груба характеристика на емпиричните факти.

C (w1 w2)	word1 – w1	word2 – w2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Таблица 1. Честота на биграми от думи в основния текст.

Първото подобрение на Таблица 1 в посока на разглежданата от нас задача е да разглеждаме биграми от думи, които са извадени от текстове обработени с POS-tagger. Тъй като той маркира частите на речта, можем да броим само честотата на двойки от значещи думи и така да наблюдаваме "сурови" колокации от по две думи. В таблица 2 са дадени шаблони за наблюдения на колокации, изразени чрез частите на речта.

Tag pattern	***Example***
Adj Noun (A N)	linear function
Noun Noun (N N)	regression coefficients
Adj Adj Noun (A A N)	Gaussian random variable
Adj Noun Noun (A N N)	cumulative distribution function
Noun Adj Noun (N A N)	mean squared error
Noun Noun Noun (N N N)	class probability function
Noun Prep Noun (N P N)	degrees of freedom

Таблица 2: образци за търсене на комбинации от две съседни части на речта.

C (w1 w2)	***W1***	***W2***	***Tag Pattern***
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gilf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N

C (w1 w2)	*W1*	*W2*	*Tag Pattern*
1073	real	estate	A N

Таблица 3. Най-чести биграми от "значещи" думи

Така ще филтрираме съчетанията от "сервизни" (незначещи) думи. Вадим всички думи, които удовлетворяват шаблоните от табл. 2 и броим честотата им. Резултатът е показан на таблица 3. Виждаме доста устойчиви словосъчетания, много имена, но и важни съставни термини като *вице-президент*. York City е остатък от 3-думово име (New York City) и в таблица 3 излиза като *често* понеже си служим с 2-местен шаблон (но на практика се започва с по-дълги шаблони и те се прилагат първи, преди кратките). Някои словосъчетания са чисто композиционни – *миналата година*, *миналата седмица* и т.н., но са забелязани имена и важни съставни термини.

На таблица 4 са извадени двадесетте най-чести сращания на двуместни колкокации със *strong* и *powerful* отляво. Пак забелязваме, че простият метод базиран на честотите е изненадващо добър. От табл. 4 можем да извадим заключение, че се употребява *strong challenge* и *powerful computer*, а не *strong computer* и powerful challenge. Но същевременно виждаме ограниченията на метода "вземи най-честите". Съществителните *man* и *force* се използват и със *strong*, и с *powerful* (макар да не е показано на табл. 4, *strong force* се появява по-надолу в стълб 1 с честота 4). За разрешаване на такива случаи има нужда от по-прецизни техники. Вижда се освен това, че *strong tea* или *powerful tea* не се срещат в разглеждания вестникарски корпус, но ако се потърси в Интернет, за първото словосъчетания могат да се извадят 799 примера, а за второто – 17 (главно в статии по лигвистика). Оттука вадим заключението, че коректния израз е *strong tea.*

Досегашните примери показват нещо важно: прости филтри с минимум лингвистично знание позволяват да се напредне значително в търсенето на словосъчетания. По-надолу ще предполагаме, че от корпуса са извлечени само думите, които се срещат най-често като **глаголи**, **съществителни** и **прилагателни** (все едно че са филтрирани само тези думи). Ще разгледаме техники за разпознаване на 2-местни колокации от такива думи.

w	*C (strong, w)*	*w*	*C (powerful, w)*
Support	50	Force	13
Safety	22	Computers	10
Sales	21	Position	8
Opposition	19	Men	8
Showing	18	Computer	8
Sense	18	Man	7
Message	15	Symbol	6
Defense	14	Military	6
Gains	13	Machines	6
Evidence	13	Country	6
Criticism	13	Weapons	5

w	*C (strong, w)*	*w*	*C (powerful, w)*
Possibility	11	Post	5
Feeling	11	People	5
Demand	11	Nation	5
Challenges	11	Forces	5
Challenge	11	Chip	5
Case	11	Germany	5
Supporter	10	Senators	4
Signal	9	Neighbor	4
Man	9	Magnet	4

Таблица 4: най-чести двуместни колкокации със *strong* и *powerful* отляво.

Таблици 3 и 4 показват добри резултати, но за съседни думи. Има обаче словосъчетания, които са устойчиви - макар думите-съставки да не са задължително непосредствени съседи. Разгледайте примерни употреби на глагола knock с един от най-честите му аргументи, door:

1.1. She knocked on his door (отдалечени на 3 думи),
1.2. They knocked at the door (отдалечени на 3 думи),
1.3. 100 women knocked on Donaldson's door (отдалечени на 5 думи – с ' и s),
1.4. A man knocked at the metal frond door (отдалечени на 5 думи).

Очевидно разстоянието между глагола и допълнението не е фиксирано, понеже може да има различен брой думи помежду им, но като че ли става дума за устойчиво съчетание и от тук ни се иска да можем да извадим заключението, че коректното изказване на английски е че на вратата се чука (knock), а не се употребяват глаголи като hit, beat и rap. Ако имаме такива техники, ще можем да извличаме автоматично знания,че решенията се "правят", а не се вземат и така нататък. Това би било много полезно.

2.2. Средно и дисперсия

Тук ще покажем как се откриват биграми от думи с фиксирана дължина на отдалеченост (а лесно се вижда как може да стане с променлива), като за простота разглеждаме контекст от 3 думи наляво и надясно.

Sentence: Stocks crash as rescue plan teeters
Bigrams: stock crash stocks as stocks rescue
crash as crash rescue crash plan
as rescue as plan as teeters
rescue plan rescue teeters
plan teeters

Фигура 1. Използване на 3-думов "колокационен прозорец" за откриване на биграми от "отдалечени" думи

На фигура 1 е показано как става това – комбинираме по двойки думите отдалечени до 3 от текущата. Получаваме биграми на думи раздалечени до 3 позиции във всички изречения на корпуса. Сега трябва да оценим кои от намерените двойки са колокации. Ще изчисляваме средната отдалеченост и дисперсията. Средната отдалеченост между глагола knock и door в примерите 1.1-1.4 е:

$$1/4\ (3+3+5+5)= 4.0$$

Ако имаше срещане на door преди knock, щеше да се преброи като отрицателно число, например като (-3) за фразата the *door that she knocked on*. Сега вече сме подготвени да "мерим" всякакви случаи.

Дисперсията показва колко индивидуалните резултати се отличават от средното. Смятаме я по формулата:

$$s^2 = \sum_{i=1, \ldots, n} (d_i-d)^2 \ / \ n$$

където *n* е броят на примерите при които две думи са срещат като колокации в нашия експеримент, d_i е отдалечеността за *i*-тия пример и *d* е средното аритметично на отдалечеността за тези две думи. Ако във всички случаи имаме еднакво отдалечени думи на биграмата, дисперсията е нула (например ако Ню Йорк е винаги заедно, дисперсията на съставящите думи за тази колокация е нула). Както е прието, ние ще използваме *отклонението* $s=\sqrt{s^2}$ (квадратен корен от дисперсията), за да добием представа за промяната на разстоянието между две думи. За knock и door в примерите 1.1-1.4 отклонението е:

$$s = \sqrt{\ 1/4\ (\ (3-4.0)^2 + (3-4.0)^2 + (5-4.0)^2 + (5-4.0)^2) = 1}$$

Средното аритметично и отклонението характеризират "сцеплението" на двойки думи в корпуса. Очевидно кандидатите за колокации трябва да имат малко отклонение, понеже то е показател доколко близко се срещат думите в целия корпус. В таблица 5 са извадени примерните *средно* и *отклонение* за двойки думи, като е работено с контекст от 9 думи наляво и надясно.

s	d	брой срещания	w1	w2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	organizations
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

Таблица 5. Намиране на колокации с използване на средното и отклонението. Примерни стойности за 12 двойки от думи.

На таблица 5 намираме видовете двойки които могат да се откриват с тази техника. Ако средното е близко до 1 и отклонението е малко, както при New York, значи имаме вид фраза която би могла да бъде открита с филтрите базирани на части на речта, както беше в таблица 2. Ако средното е доста по-голямо от 1 и отклонението е малко, имаме по-интересен случай, който няма да се разпознае с простите техники описани по-горе. Такива са:

- previous / games (с разстояние 2), която съответства на изрази като *in the previous 10 games* или *in the previous 15 games*;
- minus / points съответства на изрази *minus 2 percentage points* и *minus 3 percentage points*;
- hundreds / dollars съответства на изрази *hundreds of billions of dollars* и *hundreds of millions of dollars*.

Голямото отклонение свидетелствува, че двете думи не са в интересно за нас отношения една спрямо друга. Такива са средните 4 реда на таблица 5. Забележете, че средното е близко до нула, както бихме могли да очакваме за нормалното равномерно разпределение. Интересни са също така последните 4 реда на таблица 5. Виждаме 2 типични употреби на прилагателно-съществително и имена. В израза Richard {M.} Nixon разстоянието между собственото име и фамилията е до 4 думи (заради съкратеното втори име и точката, а Garrison said и said Garrison са смятани заедно.

Този метод за откриване на колокации с използване на средното и отклонението е предложен през 1993 и се счита за много успешен при откриване на терминология (термини от няколко думи се разпознават с точност до 80%) и при намиране на словосъчетания за приготвяне на речник, ориентиран към генерация на естествен език. При последния случай ние се интересуваме от типичните употреби в дадената област (напр. че "решенията се правят") и автоматичното намиране на "сурови" изрази и "подсказки" в корпус е голямо улеснение.

2.3. Тестване на хипотези

По принцип би следвало да се усъмним, че *високата честота* или *малкото отклонение* може да са *случайни*. Ако имаме две много чести думи, като например *new* и *company* в разглеждания корпус от вестник, то тогава *new company* може да се среща заедно достатъчен брой пъти за да ни прилича на колокация, но всъщност е нормална фраза от прилагателно и съществително (просто в този вестник за това се говори). Забележете, че още не сме говорили за условната вероятност, една дума да се среща винаги когато се среща друга дума, но все пак думите влизат в комбинации с много други думи и не можем така лесно да кажем, че някои се използват САМО като компания на определени други думи.

Стигаме до един от класическите проблеми на статистиката: да се провери над наличните данни дали едно нещо е случайно събитие или не. Ще наричаме тази проверка тестване на хипотеза. Формулираме нулевата хипотеза H_0: че няма връзка между срещането на две думи, освен случайното срещане. После ще изчислим вероятността p тези думи да се срещнат ако H_0 е истина и след това:

- ще "отречем" H_0 ако p е твърде малка (например при стойност $p<0.05$, което обикновено се приема в експерименталните науки, или друга стойност която фиксираме след разглеждане данните на експеримента) или

- ще приемем H_0 за вярна.

Както по-горе, ще гледаме за съвместни срещания на биграми в данните, но вече ще вземаме предвид колко данни сме видели. Дори да имаме много често повтаряне на дадена двойка от думи, няма да го приемаме за колокация, ако не сме наблюдавали достатъчно данни – понеже при малко данни това може да е случайно.

Първо формулираме нулевата хипотеза, какво трябва да е изпълнено за двойка думи, които не формират колокация. Тогава те са свободна комбинация от думи и са генерирани заедно напълно независимо една от друга, тоест шансът за тяхното появяване заедно е вероятността:

$$P(w1\ w2) = P(w1)\ P(w2)$$

Тоест произведение от вероятностите за поява на отделните думи. Това е прост модел, който не е потвърден емпирично, но засега ни служи като нулева хипотеза.

2.3.1. *t*-тест

Сега се нуждаем от статистически тест, който да ни казва колко е вероятно или невероятно да се случи дадена конфигурация. Един широко използван тест при задачата за търсене на колокации е *t*-теста. При него се гледат средното и отклонението на примерна извадка от случаи, където нулевата хипотеза е, че примерът е изваден от нормално разпределение със средно μ. За да изчислим вероятността да се подбере нашата примерна извадка, пресмятаме т. нар. *t*-статистика:

$$t = (x - \mu) / \sqrt{(s^2/N)},$$

където x е средното на извадката, s^2 е нейната дисперсия, N е размерът ѝ, и μ е средното на разпределението. Ако тази *t*-статистика е достатъчно голяма, можем да отречем нулевата хипотеза. Таблицата на *t*-разпределнието може да се види в статистически справочници.

Сега ще дадем прост пример как се прилага *t*-теста. Нека нашата нулева хипотеза е, че средната височина (на една извадка) на хората е 158 см. Дадени са ни примери за 200 души със средна височина 169 см (x=169) и дисперсия s^2=2600 и искаме да установим дали тази извадка е от типичното население (това ни е нулевата хипотеза) или пък е от дадено подмножество от по-ниски хора. По горната формула получаваме:

$$t = (169 - 158) / \sqrt{(2600/200)} = 3.05 \text{ (приблизително)}$$

Като погледнем в статистическите таблици за *t*-разпределнието, виждаме че стойността която съответства на ниво на достоверност 0.005 е t=2.576. Той като сме получили по-голямо t, а именно 3.05, то можем да отхвърлим нулевата хипотеза с много голяма вероятност (99.5%) и вероятността за грешка е 0.5%. И така на практика излиза, че сме взели извадка, която не е от типични хора.

Сега ще прилагаме теста за колокации. Изчисляваме над наличните ни данни *t* за *new companies*. Над коя извадка да измерим средното и дисперсията? Ще работим с целия корпус и всички биграми от този вид. Първо пресмятаме вероятността да се случат думите *new* и *companies,* които се срещат съответно 15828 и 4675 пъти в корпус от 14307668 словоформи:

$$P(new) = 15828/14307668 \qquad P(companies) = 4675/14307668$$

И после пресмятаме вероятността независимите събития да се случат заедно:

Нулева хипотеза е че двете думи се срещат заедно като независими с вероятност:

$H_{0:}$ P(*new companies*) = P(*new*) . P(*companies*) = 15828/14307668 . 4675/14307668 = $3.615 . 10^{-7}$

Ако нулевата хипотеза е вярна, двете думи са генерирани случайно една след друга, и това е ефект на един опит на Бернули с вероятност 3.615×10^{-7}. (Все едно сме хвърляли два зара с по 14307668 страни и двете думи са се паднали заедно). За такова разпределение знаем, че $\mu=3.615 \times 10^{-7}$ и дисперсията е $\sigma^2 = p(1-p)$. Нека апроксимираме отклонението σ с *p,* поради малката вероятност на повечето биграми. Сега имаме всички данни да пресметнем стойността на t, за да направим оценка на достоверността по t-теста. Оказва се, че има 8 срещания на *new companies* в корпуса от 14 млн и повече думи, така че $x=8/14307668=5.59110^{-7}$ и тогава

$$t = (x - \mu) / \sqrt{(s^2/N)} = (5.59110^{-7} - 3.615 . 10^{-7}) / \sqrt{(5.59110^{-7}/14307668)} = 0.999932.$$

По този начин t-стойността не е по-голяма от цитираната по-рано 2.576, което е критичната стойност за достоверност 0.005. И така не можем да отхвърлим нулевата хипотеза, че двете думи се срещат една до друга случайно и не формират колокация. Значи я приемаме за вярна, което в този случай изглежда прекрасно, понеже става дума за обикновено съчетание на прилагателно и съществително с композиционно значение. И така, като вадим биграми, можем да проверяваме с дадена степен на достоверност дали те формират колокации.

Таблица 6 показава t-стойностите за 10 биграми, които се случват точно по 20 пъти в текста. За горните 5 биграми, можем да отхвърлим нулевата хипотеза за случайно срещане с достоверност 0.005, така че те са добри кандидати да бъдат обявени за "колокации". Долните 5 биграми обаче не минават през теста за достоверност и ги отхвърляме, за нас те няма да влязат в списъка на намерените колокации.

t	C(w1)	C(w2)	C(w1 w2)	w1	w2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

Таблица 6. Прилагане на t-теста към 10 биграми, които се срещат точно по 20 пъти

Забележете, че с използване само на честотата не бихме могли да вземем решение за тези биграми, понеже те всичките се срещат по 20 пъти (поравно) и биха изглеждали равновероятни. На таблица 6 се вижда, че t-тестът взема честотата на съвместното срещане C(w1 w2) и я обработва заедно с честотите на отделните думи C(w1) и C(w2). При много често срещане на отделните две думи и малко срещане заедно, не очакваме

да имаме колокация - а случайно комбиниране. Интуитивно това има смисъл, както и обратното (по-рядко срещане, предимно заедно).

Да коментираме и факта, че на таблица 6 има стоп-думи включени в колокациите. Оказва се, че езикът е доста регулярен по отношение на употребяваните думи в даден жанр. Има 831 биграми в този корпус, които се срещат точно по 20 пъти, и за 824 от тях нулевата хипотеза може да се отхвърли. Причината за това голямо число е фактът, че има съвсем малък брой нерегулярни и случайни (непредсказуеми) явления. Така че статистическите методи дават добри резултати и има надежда за успех на подходите за разпознаване на колкокациите и word-sense disambiguation (които са още в доста начална фаза на развитие).