Задача 1. В таблица 1 е дадена колекция от два документа d_1 и d_2 и въпрос q, постъпващ в търсачка за извличане на документи. (За естествения интелект е ясно, че само d_2 е релевантен на q, изкуствения обаче трябва да го сметне). Ще се упражним в мини-търсене само спрямо значещите думи от въпроса.

Документ d_1	Коси слънчеви лъчи осветиха русите коси на Мария. Няколко коса накацаха до прозореца и един запя. Мария надникна изпод разпилените си коси. Песента на коса я разсъни, тя се изми и среса дългата си коса.
Документ d_2	Бащата на Мария взе коса и тръгна да коси ливадата. Той работи с много остра коса. Брат й също обича да коси, но днес не отиде, понеже не намериха още една коса.
Въпрос q	Кой коси с коса? (Забележка: *кой* и *с* са стоп-думи)

Таблица 1. Документи и въпрос в необработен (суров) вид

Задание 1. В таблици 2, 3 и 4 препишете по-важните за нас думи от d_1, d_2 и q така, както те биха изглеждали ако системата извършваше над тях съответно: *stemming* по последна буква, морфологичен анализ+*POS-tagging* и морфологичен анализ + *POS-tagging* + *word-sense disambiguation*

Документ $d_{1\text{-}1}$	
Документ $d_{2\text{-}1}$	
Въпрос q_1	

Таблица 2. След прилагане на stemming спрямо последната буква

Документ $d_{1\text{-}2}$	
Документ $d_{2\text{-}2}$	
Въпрос q_2	

Таблица 3. След прилагане на морфологичен анализ+POS-tagging

Документ $d_{1\text{-}3}$	
Документ $d_{2\text{-}3}$	
Въпрос q_3	

Табл.4. След прилагане на морфологичен анализ+POS-tagging+ word-sense disambiguation (WSD)

Задание 2. Очевидно системата строи модел на векторното пространство спрямо тези думи, които е способна да разпознае във въпроса и намери след това в документите. При обработка на сурови текстове както в табл. 1, тя би си послужила с оси от вида КОСИ и КОСА за значещите низове от въпроса. При способности за по-фин анализ, тя би работила спрямо пространството КОСЯ и КОСА (това съответства на табл. 3 и 4), но би броила различни неща в двата случая. При stemming би използвала само една ос КОС* (табл. 2).

Спрямо пространството от низовете КОСИ/КОСА за таблица 1, едномерното КОС* за таблица 2 и спрямо думите КОСЯ/КОСА за таблици 3 и 4, направете:

а) постройте само по честотите на разглежданите думи вектори за d_1, d_2 и q в съответните едно- и двумерни пространства с нормализация така, че

$$\sqrt{\sum_{i=1,\ldots,n} d_i^2 = 1}$$

Начертайте векторите схематично. Не задълбавайте много в сметките, тъй като относителното разположение на векторите се вижда с просто око.

б) пресметнете косинус-близостта между векторите и въпроса в 4-те случая, като ползвате следната формула за разстояние между въпрос и документ:

$$\cos(\mathbb{q},\mathbb{d}) = \sum_{i=1,\ldots,n} q_i \cdot d_i \;/\; \sqrt{\sum_{i=1,\ldots,n} q_i^2} \quad \sqrt{\sum_{i=1,\ldots,n} d_i^2}$$

където $\mathbb{q}$ и $\mathbb{d}$ са векторите на въпроса и документа. Грубите сметки показват кой е най-малкият ъгъл между вектора-въпрос и векторите-документи.

в) Представете си, че системата трябва да извлече само един документ – този с най-голяма релевантност (или повече документи, ако са еднакво подобни на въпроса). Попълнете долната таблица, за да отговорите: кой документ е най-близо до въпроса, тоест при кой вид обработка на текста получаваме най-релевантен резултат по метода на векторното пространство?

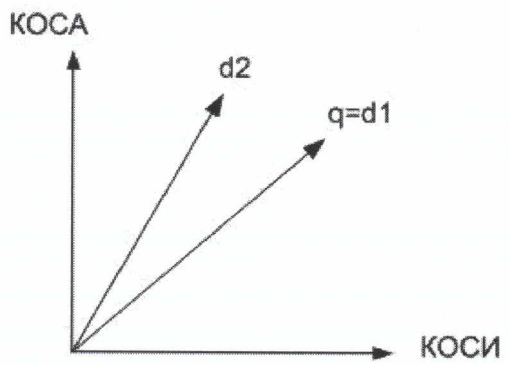
	В суров вид, Таблица 1	След stemming, Таблица 2	След морфолог. анализ и POS-tagging, Табл. 3	След морфолог. Анализ, POS-tagging и WSD, Табл. 4
Най-близо до въпроса е:				

Решение задача 1. Щом разглеждаме мини-търсене само спрямо значещите думи от въпроса, то винаги първо определяме как изглеждат въпросът и документите. Тоест, в четирите дадени случая на обработка, търсещата система работи с различни въпроси и документи поради своите различни способности да разпознае значещите думи. Нека сега вникнем в разликите за всеки случай поотделно.

Документ d_1	Коси ..... коси ..... коса ..... коси ..... коса ..... коса
Документ d_2	..... коса ..... коси ..... коса ..... коси.....коса.
Въпрос q	коси коса

Таблица 1. Документи и въпрос в необработен (суров) вид – разглеждаме само низовете КОСИ и КОСА. Търсещата програма не може да разпознава нищо друго освен низове, като е способна единствено да изтрие стоп-думите и нерелевантните думи

Разглеждаме само думите от въпроса. На практика това значи, че в многомерното оригинално пространство (на всички думи от всички документи) се фокусираме само върху една равнина определена от две оси, за двете значещи думи. Във въпроса q низовете КОСИ и КОСА се срещат по един път. Векторът на въпроса е единичен под ъгъл 45 градуса към абцисата/ординатата. В d_1 КОСИ се среща 3 пъти, а КОСА – пак 3 пъти. Векторът на d_1 съвпада с вектора q след нормиране. В d_2 КОСИ се среща 2 пъти, а КОСА –3 пъти. Векторът на d_2 е по-близо до оста КОСА отколкото до КОСИ, в съотношение 3:2. Значи във финалната таблица записваме d_1 - тъй като той е най-близкият документ до въпроса (тоест най-близкият вектор до вектора на въпроса). Така че за този случай имаме следната картинка от вектори и запълнена таблица:



	В суров вид, Таблица 1	След stemming, Таблица 2	След морфолог. анализ и POS-tagging, Табл. 3	След морфолог. Анализ, POS-tagging и WSD, Табл. 4
Най-близо до въпроса е:	d_1			

Минаваме към втория случай. При stemming имаме едномерно пространство спрямо оста КОС*. При нормиране всички вектори съвпадат с единичния. Спрямо интересуващата ни основа КОС*, и двата документа са еднакво релевантни. Записваме го в таблицата.

Документ d_1	Коси слънчеви лъчи осветиха русите коси на Мария. Няколко коса накацаха до прозореца и един запя. Мария надникна изпод разпилените си коси. Песента на коса я разсъни, тя се измъкна и среса дългата си коса. КОС* - 6 пъти
Документ d_2	Бащата на Мария взе коса и тръгна да коси ливадата. Той работи с много остра коса. Брат й също обича да коси, но днес не отиде, понеже не намериха още една коса. КОС* - 5 пъти
Въпрос q_1	КОС* - 2 пъти

Таблица 2. След прилагане на stemming спрямо последната буква

	В суров вид, Таблица 1	След stemming, Таблица 2	След морфолог. анализ и POS-tagging, Табл. 3	След морфолог. Анализ, POS-tagging и WSD, Табл. 4
Най-близо до въпроса е:	d_1	d_1, d_2		

Минаваме към третия случай. Системата извършва морфологичен анализ и разрешава многозначността на частите на речта. Така тя разпознава във въпроса един глагол и едно съществително, после търси именно тях в документите. При съществителното КОСА в d_1 се смесват формите на две съществителни – КОСА и КОС, но процесът на POS-tagging най-често не може да разпознае това смесване, тъй като той ще определи и двете форми като „съществително". Той ще покаже обаче, че първата дума в d_1 е форма на прилагателно – понеже след нея стои друго прилагателно – и така първата дума КОСИ в d_1 няма да се брои към търсената конфигурация. В този случай документът d_2 е по-близо до въпроса.

Документ d_1	Коси слънчеви лъчи осветиха русите коси на Мария. Няколко коса накацаха до прозореца и един запя. Мария надникна изпод разпилените си коси. Песента на коса я разсъни, тя се измъкна и среса дългата си коса. КОСА съществително - 5 пъти Глагол КОСЯ – 0 пъти
Документ d_2	Бащата на Мария взе коса и тръгна да коси ливадата. Той работи с много остра коса. Брат й също обича да коси, но днес не отиде, понеже не намериха още една коса. КОСА съществително - 3 пъти Глагол КОСЯ – 2 пъти
Въпрос q_1	КОСЯ глагол 1 път, КОСА съществително 1 път

Таблица 3. След прилагане на морфологичен анализ+POS-tagging

Таблица с отговорите след случай 3:

	В суров вид, Таблица 1	След stemming, Таблица 2	След морфолог. анализ и POS-tagging, Табл. 3	След морфолог. Анализ, POS-tagging и WSD, Табл. 4
Най-близо до въпроса е:	d_1	d_1, d_2	d_2	

Минаваме към случай 4. Системата е безкрайно интелигентна и прави всичко, което можем да си представим. Очевидно d_1 не е подобен на въпроса, понеже не съдържа нито една от значещите думи, разпознати във въпроса.

Документ d_1	Коси слънчеви лъчи осветиха русите коси на Мария. Няколко коса накацаха до прозореца и един запя. Мария надникна изпод разпилените си коси. Песента на коса я разсъни, тя се изми и среса дългата си коса. КОСА съществително селско-стопански уред - 0 пъти Глагол КОСЯ – 0 пъти
Документ d_2	Бащата на Мария взе коса и тръгна да коси ливадата. Той работи с много остра коса. Брат й също обича да коси, но днес не отиде, понеже не намериха още една коса. КОСА съществително селско-стопански уред - 3 пъти Глагол КОСЯ – 2 пъти
Въпрос q_1	КОСЯ глагол 1 път, КОСА съществително селско-стопански уред 1 път

Табл.4. След прилагане на морфологичен анализ+POS-tagging+ word-sense disambiguation (WSD)

Така таблица с отговорите на задача 1 е:

	В суров вид, Таблица 1	След stemming, Таблица 2	След морфолог. анализ и POS-tagging, Табл. 3	След морфолог. Анализ, POS-tagging и WSD, Табл. 4
Най-близо до въпроса е:	d_1	d_1, d_2	d_2	d_2