**KAUNO TECHNOLOGIJOS UNIVERSITETAS**

**INFORMATIKOS FAKULTETAS**

DIRBTINĖS INTELEKTIKOS PAGRINDAI

Laboratornis darbas Nr. 3 – Bajeso teorema

Atliko:

Laimonas Sutkus IFF-4/1

**KAUNAS, 2017**

**Darbo užduotis:**

Parašyti spam filtrą naudojant Bajeso teoremą.

**Programos kodas:**

|  |
| --- |
| **from** \_\_future\_\_ **import** division **from** collections **import** Counter **import** os **import** re **import** sys **import** pickle  **if** len(sys.argv) < 5:  **raise** EnvironmentError(**"1) path to non-spam dir 2) path to spam dir 3) path to a file to "  "check 4) number of words to consider 5) ignore extremums"**)  *# path to directory which contains files that are marked as not spam* pathToNotSpam = sys.argv[1]  *# path to directory which contains files that are marked as spam* pathToSpam = sys.argv[2]  *# path to a file to be checked if it is spam or not* pathToFileValidation = sys.argv[3]  *# path to serialized word table with spam probabilities (path or "-")* pathToSerializedTable = sys.argv[4]  *# number of words to consider when calculating files "spaminess"* numberOfWords = sys.argv[5] numberOfWords = 150  *# ignore words that have spam rating of 0 percent or 100 percent (y, n)* ignoreExtremum = True **if** sys.argv[6] == **'y' else** False   **def** parseFile(pathToFile):  lines = open(pathToFile, **"r"**).read().splitlines()  words = filter(**lambda** s: s, [filter(**lambda** s: s != **""**, re.split(**"[^a-zA-Z0-9\']"**, line)) **for** line **in** lines])  **return** words  **if** pathToSerializedTable == **"-"**:  **def** getCounts(pathToFiles):  wordTable = []  **for** fileName **in** os.listdir(pathToFiles):  wordTable.append(parseFile(pathToFiles + fileName))  wordTable = Counter(sum([sum(lines, []) **for** lines **in** wordTable], []))  **return** wordTable    spamTable = getCounts(pathToSpam)  nonSpamTable = getCounts(pathToNotSpam)   allWords = list(set(spamTable.keys()) | set(nonSpamTable.keys()))   pSpamTable = {key: value / len(allWords) **for** key, value **in** {key: spamTable[key] **for** key **in** allWords}.iteritems()}  pNonSpamTable = {key: value / len(allWords) **for** key, value **in** {key: nonSpamTable[key] **for** key **in** allWords}.iteritems()}   probOfSpam = {key: pSpamTable[key] / (pSpamTable[key] + pNonSpamTable[key]) **for** key **in** allWords}  pickle.dump(probOfSpam, open(**"./wordtable.p"**, **"wb"**)) **else**:  probOfSpam = pickle.load(open(**"./"** + pathToSerializedTable, **"rb"**))  checkedFileWords = sum(parseFile(pathToFileValidation), []) checkedFileSpamProb = [probOfSpam[word] **if** word **in** probOfSpam **else** 0.40 **for** word **in** checkedFileWords] checkedFileSpamProb.sort(reverse=True)  **if** ignoreExtremum:  checkedFileSpamProb = [prob **for** prob **in** checkedFileSpamProb **if** prob != 0.0 **and** prob != 1.0][0:int(numberOfWords)] **else**:  checkedFileSpamProb = checkedFileSpamProb[0:numberOfWords]  oppositeCheckedFileSpamProb = [1 - self **for** self **in** checkedFileSpamProb]  spamProbabilityMultiplied = reduce(**lambda** x, y: x \* y, checkedFileSpamProb) oppositeCheckedFileSpamProbMultiplied = reduce(**lambda** x, y: x \* y, oppositeCheckedFileSpamProb)  spamProbability = spamProbabilityMultiplied / (spamProbabilityMultiplied + oppositeCheckedFileSpamProbMultiplied)   **print**(spamProbability) |

**Filtro testavimas:**

|  |
| --- |
| **Ne spam laiškas (134 žodžiai)** |
| [Razor-users] Re: Can't use and undefined value.." error> Date: Mon, 30 Sep 2002 14:00:12 -0400 (EDT)  > From: Dayv Gastonguay <noghri@nauticom.net>  >  > I just installed razor 2.152 on a FreeBSD 4.4-RELEASE box and having  > problems with razor-check. Any time razor-check is run, (with or without  > arguments), i get this error:  >  > Can't use an undefined value as a symbol reference at  > /usr/local/lib/perl5/site\_perl/5.005/i386-freebsd/Razor2/Client/Agent.pm  > line 756.  >  Try installing the latest Perl (at least 5.6.1) port on Freebsd and make  sure you set the system to use perl from ports (i.e. in the  ports/lang/perl5/files directory run ./use.perl port. Reinstall the relevant  perl modules needed by razor and try again.  Sven |
| **Spamiškumo tikimybė (125 žodžių) - 4.69071124245e-67** |
| **Spamiškumo tikimybė (100 žodžių) - 1.46171562083e-39** |
| **Spamiškumo tikimybė (50 žodžių) - 2.13575174648e-07** |
| **Spamiškumo tikimybė (25 žodžių) - 0.249729575684** |
| **Spamiškumo tikimybė (15 žodžių) - 0.92029361313** |

|  |
| --- |
| **Spam laiškas (73 žodžiai)** |
| [SPAM] Diamonds on your watch will tell a lot about you.From nobody Sun Jan 03 19:33:40 2016  Content-Type: text/plain;  charset="Windows-1252"  Content-Transfer-Encoding: quoted-printable  Visit us right now to change your life status up to a successful person. Th=  e irreplaceable accessory of every successful person is an expensive watch.=  You can get copies of original Swiss brands at our site!  &nbsp;  A smart click |
| **Spamiškumo tikimybė (70 žodžių) - 0.0040372335264** |
| **Spamiškumo tikimybė (50 žodžių) - 0.999998388938** |
| **Spamiškumo tikimybė (25 žodžių) - 0.999999999999** |
| **Spamiškumo tikimybė (15 žodžių) - 1.0** |
| **Spamiškumo tikimybė (10 žodžių) - 1.0** |

|  |
| --- |
| **Maišytas laiškas (207 žodžiai)** |
| [SPAM] Diamonds on your watch will tell a lot about you.From nobody Sun Jan 03 19:33:40 2016  Content-Type: text/plain;  charset="Windows-1252"  Content-Transfer-Encoding: quoted-printable  Visit us right now to change your life status up to a successful person. Th=  e irreplaceable accessory of every successful person is an expensive watch.=  You can get copies of original Swiss brands at our site!  &nbsp;  A smart click  [Razor-users] Re: Can't use and undefined value.." error> Date: Mon, 30 Sep 2002 14:00:12 -0400 (EDT)  > From: Dayv Gastonguay <noghri@nauticom.net>  >  > I just installed razor 2.152 on a FreeBSD 4.4-RELEASE box and having  > problems with razor-check. Any time razor-check is run, (with or without  > arguments), i get this error:  >  > Can't use an undefined value as a symbol reference at  > /usr/local/lib/perl5/site\_perl/5.005/i386-freebsd/Razor2/Client/Agent.pm  > line 756.  >  Try installing the latest Perl (at least 5.6.1) port on Freebsd and make  sure you set the system to use perl from ports (i.e. in the  ports/lang/perl5/files directory run ./use.perl port. Reinstall the relevant  perl modules needed by razor and try again.  Sven |
| **Spamiškumo tikimybė (200 žodžių) - 1.90142617055e-69** |
| **Spamiškumo tikimybė (150 žodžių) - 3.05838437354e-23** |
| **Spamiškumo tikimybė (100 žodžių) - 0.254628693271** |
| **Spamiškumo tikimybė (50 žodžių) - 0.99999999912** |
| **Spamiškumo tikimybė (10 žodžių) - 0.999999999627** |

**Išvados:**

Spam filtro tikslumas priklauso nuo laiško žodžių skaičiaus ir testinių žodžių skaičiaus. Kuo didesnis laiškas, tuo tikslingiau paduoti daugiau testinių žodžių iš laiško.