Inngangur að máltækni - Verkefni 5

Guðmundur Óli Norland

1

Ég notaðist einungis við *mbl.txt* skránna í Gullstaðlinum.

a)

Nefnir virðist lemma töluvert betur en Lemmald. Nefnir er án efa betur þjálfaður á meðan Lemmald reynir að lemma út frá endingu og marki sem virðist mistakast nokkuð reglulega. Í kóðanum prenta ég dæmi um vitleysur. Dæmi sem ég tók fyrir Lemmald eru *('undirtitilinn', 'nkeog', 'undirtitilur')* og *('Áfangar', 'nkfn-s', 'áfangur')*. Það er erfitt að finna villur hjá Nefni en þó á hann í erfiðleikum með slangur og talmál. Dæmi sem ég tók fyrir nefni er *('soldil', 'aa', 'soldil')* þar sem lemman ætti að vera *svolítill*.

b)

Hér fyrir neðan í úttakshlutanum má sjá nánar dæmin sem ég vitna í. Fyrsta dæmið sem ég fann þar sem Lemmald og Nefnir lemma ólíkt er að Nefnir breytir ekki kr. á meðan Lemmald lemmar kr. sem króna. Þegar ég skoðaði hlutfall lesmálsorða sem eru lemmuð ólíkt fékk ég 3.6%, niðurstöðuna fékk ég með því að zipa saman listana, ítra svo og telja tilvikin þar sem lemmunin er ólík. Dæmi um eins lemmun en hvort tveggja vitlaust, þá fann ég á einum stað að orðmyndin *hægt* (*...leikir fara fram er hægt að...*) er eins lemmuð sem *hægur* sem er röng niðurstaða í því samhengi þar sem orðið er rangt markað. Í því samhengi er þetta atviksorð en er markað sem lýsingarorð. Dæmi um mismunandi lemmun og hvort tveggja vitlaust þá fann ég á einum stað að orðmyndin *fyrst* er rangt lemmuð í báðum tilvikum, ætti að vera *snemma*.

2

a)

Þetta verkefni framkvæmi ég í fallinu *ver2a*. Fyrst les ég í gegnum skránna *"bin_mim_transformation.txt"* og geymi gögnin í lista. Því næst les ég í gegnum skránna *"SHsnid.csv"* þar sem fyrir hverja línu leita ég í BÍN-MÍM listanum að marki með sama flokk og greiningarstreng og er í línunni og bæti svo við í nýjan lista. Út frá þessum nýja lista er svo nýja skráin búin til (*"nytt_snid.txt"*).

b)

Þetta verkefni framkvæmi ég í fallinu *ver2b*. Ég byrja á að geyma línurnar í nýja skjalinu úr síðasta lið (*"nytt_snid.txt"*) í lista sem ég svo leita í þegar ég ítra í gegnum línurnar úr gullstaðlinum, ef ekkert finnst fæst *UNK*. Niðurstöðurnar eru svo geymdar í nýrri skrá (*"min_lemmun.txt"*).

c)

Þetta verkefni framkvæmi ég í fallinu *ver2c*. Til að bera saman mína lemmun og Nefni / Lemmald þá zipa ég saman listana, ítra svo í gegnum þá og tel ólík stök (og deili svo með heildarfjöldanum). Fyrir nefni fékk ég 32.4% ólíkt en fyrir Lemmald 34.1%. Hlutfall *UNK* er 28.7%.

Kóði

verkefni5.py

```
import nefnir
import os
import time

mappa = os.path.dirname(os.path.abspath(__file__))


def prosenta(tala):
    return str(round(tala * 100, 1)) + "%"


### UNDIRBÚNINGUR ###
# ----------------- #

# lesum inn gullskrá og geymum í lista á sniðinu [(orðmynd, mark)....]
def lesa_gull(filepath):
    tvenndir = []
    texti = open(filepath, "r")
    for line in texti:
        split = line.split("\t")
        if len(split) > 1:
            # orð
            o = split[0]
            # mark
            mark = split[1].rstrip()
            tvenndir.append((o, mark))
    return tvenndir


# undirbúum skrár sem við getum lemmað með IceNLP-1.4.2/IceNLP/bat/lemmald/lemmatize.sh skriptunni
def lemmald_fyrir(filepath):
    # lemmald þarf bil sem aðskilnaðartákn á meðan gullstaðallinn notar tab
    with open(filepath, "r") as f:
        data = f.read().replace("\t", " ")
        # vistum í nýrri skrá
        lemmald_fyrir_path = os.path.join(mappa, "lemmald_fyrir.txt")
        with open(lemmald_fyrir_path, "w") as lemmald_fyrir:
            lemmald_fyrir.write(data)


### LEMMUN ###
# ---------- #

# áður en þetta er keyrt þarf að vera búið að keyra lemmald_fyrir fallið og keyra lemmatize.sh
def lemmald():
    lemmald_eftir_path = os.path.join(mappa, "lemmald_eftir.txt")
    triples = []
    with open(lemmald_eftir_path, "r") as lemmald_eftir:
        for line in lemmald_eftir:
            line = line.rstrip()
            # print(line, len(line))
            if len(line) == 0:
                continue
            split = line.split(" ")
```

```
            word = split[0]
            lemma = split[1]
            mark = split[2]
            triples.append((word, mark, lemma))
    return triples


def nefnir_lemmun(tvenndir):
    triples = []
    for t in tvenndir:
        x = nefnir.lemmatize(t[0], t[1])
        triples.append((t[0], t[1], x))
    return triples


# ============================================== #
# ---------------------------------------------- #
### VERKEFNI / KEYRSLA ### --------------------- #
# ---------------------------------------------- #
# ============================================== #

# global breytur
lem = []
nefn = []


def ver1():
    gull_mbl_path = os.path.join(mappa, "MIM-GOLD-1_0/MIM-GOLD-1_0/mbl.txt")
    mbl = lesa_gull(gull_mbl_path)
    # lemmald_fyrir(gull_mbl_path)

    # fáum lista af þrenndum á forminu [(token, mark, lemma)...]
    global lem, nefn
    lem = lemmald()
    nefn = nefnir_lemmun(mbl)

    # a)
    # ---------------------------------------------
    print("Lemmald, dæmi um tvær vitleysur:")
    print(lem[2650:2660])
    print()
    print(lem[5910:5920])
    print("\nNefnir, dæmi um vitleysu:")
    print(nefn[6580:6590])

    # b)
    # ---------------------------------------------

    # sækjum allar orðmyndir sem eru lemmaðar ólíkt
    z = list(zip(lem, nefn))
    olik_lemmun = []
    for l, n in z:
        if not l[2].lower() == n[2].lower():
            olik_lemmun.append((l, n))
    # olik_lemmun = list(set(olik_lemmun))
```

```
    # sýnum dæmi um ólíka lemmun
    print("Ólík lemmun:")
    print("Við sjáum að Nefnir breytir ekki kr. á meðan Lemmald lemmar kr. sem króna")
    print("------------")
    for l, n in z[13566:13576]:
        print("Lemmald:", l)
        print("Nefnir:", n)
    print()

    # sýnum hlutföll
    print("Hlutföll:")
    print("---------")
    lemmad_olikt_hlutfall = len(olik_lemmun) / len(z)
    print(
        "Hlutfall lesmálsorða sem eru lemmuð eins:", prosenta(1 - lemmad_olikt_hlutfall)
    )
    print("Hlutfall lesmálsorða sem eru lemmuð ólíkt:", prosenta(lemmad_olikt_hlutfall))
    print()

    print("Eins lemmun en hvort tveggja vitlaust:")
    print("Orðmyndin 'hægt' er eins lemmuð en við fáum ranga niðurstöðu þar sem orðið")
    print(
        "er rangt markað. Í þessu samhengi er þetta atviksorð en er markað sem lýsingarorð."
    )
    print("-------------------------------------")
    for l, n in z[600:610]:
        print("Lemmald:", l)
        print("Nefnir:", n)
    print()

    print("Mismunandi lemmun og hvort tveggja vitlaust:")
    print("Orðmyndin 'fyrst' er rangt lemmuð í báðum tilvikum, ætti að vera 'snemma'")
    print("-------------------------------------")
    for l, n in z[248090:248095]:
        print("Lemmald:", l)
        print("Nefnir:", n)
    print()


def ver2a():
    # lesum bín transformation skránna og geymum þrenndir í fylki
    # á forminu [(orðflokkur, greiningarstrengur, mark)...]
    binmim = []
    with open(os.path.join(mappa, "bin_mim_transformation.txt"), "r") as binmim_skra:
        for line in binmim_skra:
            split = line.split(">")
            flokkur = split[0]
            gr_strengur = split[1].split("|")[0]
            mark = split[1].split("|")[1].rstrip()
            binmim.append((flokkur, gr_strengur, mark))

    # lesum næst SHsnid.csv og geymum línur á nýja sniðinu (orðmynd, mark, lemma) í lista
    nytt_snid = []
    with open(os.path.join(mappa, "SHsnid.csv"), "r") as shsnid:
        for line in shsnid:
            split = line.split(";")
```

```
            # notum orðflokkinn og greiningarstrenginn til að reyna að finna mím-mark
            flokkur = split[2]
            gr_strengur = split[5].rstrip()
            # leitum, ef finnst ekki notum við undirstrik
            mim_mark = [x[2] for x in binmim if x[0] == flokkur and x[1] == gr_strengur]
            if len(mim_mark) > 0:
                mim_mark = mim_mark[0]
            else:
                mim_mark = "_"
            # bætum línunni við í listann
            ordmynd = split[4]
            lemma = split[0]
            nytt_snid.append((ordmynd, mim_mark, lemma))
    # búum að lokum til nýja skrá með öllum línunum á nýja sniðinu
    with open(os.path.join(mappa, "nytt_snid.txt"), "w") as nytt_snid_skra:
        for line in nytt_snid:
            nytt_snid_skra.write(";".join(line) + "\n")


def ver2b(input_file):
    # geymum línurnar á nýja sniðinu í lista sem við getum leitað í
    nytt_snid = []
    nytt_snid_dict = dict()
    with open(os.path.join(mappa, "nytt_snid.txt"), "r") as nytt_snid_skra:
        i = 0
        for line in nytt_snid_skra:
            split = line.split(";")
            ordmynd = split[0]
            mark = split[1]
            lemma = split[2].rstrip()
            nytt_snid.append((ordmynd, mark, lemma))
            # geymum líka index fyrir allar orðmyndir í dict svo við getum síað listann niður
            # og þar af leiðandi leitað hratt í honum (í staðinn fyrir að ítra í gegnum hann allan)
            if ordmynd not in nytt_snid_dict:
                nytt_snid_dict[ordmynd] = [i]
            else:
                nytt_snid_dict[ordmynd] += [i]
            i += 1
    # print(len(nytt_snid))
    # lemmum input skránna
    min_lemmun = []
    gull = lesa_gull(input_file)
    for line in gull:
        ordmynd = line[0]
        mark = line[1]
        # leitum, ef ekkert finnst mun UNK haldast
        lemma = "UNK"
        if ordmynd in nytt_snid_dict:
            # síum niður listann bara fyrir þessa orðmynd
            filtera = [nytt_snid[i] for i in nytt_snid_dict[ordmynd]]
            # leitum að lemmu með þessari orðmynd og þessu marki
            lemma = [x[2] for x in filtera if x[0] == ordmynd and x[1] == mark]
            if lemma:
                lemma = lemma[0]
            else:
                # leitum að lemmu bara með þessari orðmynd
```

```
                lemma = [x[2] for x in filtera if x[0] == ordmynd and x[1] == "_"]
                if lemma:
                    lemma = lemma[0]
                else:
                    lemma = "UNK"
            # bætum við í listann
            min_lemmun.append((ordmynd, mark, lemma))

    # vistum nýja skrá með lemmuninni
    with open(os.path.join(mappa, "min_lemmun.txt"), "w") as min_lemmun_skra:
        for line in min_lemmun:
            min_lemmun_skra.write(" ".join(line) + "\n")


def ver2c():
    # keyrum verkefni 2b einu sinni til að búa til lemmaða skrá
    # ver2b(os.path.join(mappa, "MIM-GOLD-1_0/MIM-GOLD-1_0/mbl.txt"))

    # lesum svo skránna og berum saman við nefni og lemmald
    min_lemmun = []
    with open(os.path.join(mappa, "min_lemmun.txt"), "r") as min_lemmun_skra:
        for line in min_lemmun_skra:
            split = line.split(" ")
            ordmynd = split[0]
            mark = split[1]
            lemma = split[2].rstrip()
            min_lemmun.append((ordmynd, mark, lemma))
    # nefnir
    z = list(zip(min_lemmun, nefn))
    olik_lemmun = []
    for l, n in z:
        if not l[2].lower() == n[2].lower():
            olik_lemmun.append((l, n))
    lemmad_olikt_hlutfall = len(olik_lemmun) / len(z)
    print("Mín lemmun og nefnir:")
    print(
        "Hlutfall lesmálsorða sem eru lemmuð eins:", prosenta(1 - lemmad_olikt_hlutfall)
    )
    print("Hlutfall lesmálsorða sem eru lemmuð ólíkt:", prosenta(lemmad_olikt_hlutfall))
    # lemmald
    z = list(zip(min_lemmun, lem))
    olik_lemmun = []
    for l, n in z:
        if not l[2].lower() == n[2].lower():
            olik_lemmun.append((l, n))
    lemmad_olikt_hlutfall = len(olik_lemmun) / len(z)
    print("\nMín lemmun og lemmald:")
    print(
        "Hlutfall lesmálsorða sem eru lemmuð eins:", prosenta(1 - lemmad_olikt_hlutfall)
    )
    print("Hlutfall lesmálsorða sem eru lemmuð ólíkt:", prosenta(lemmad_olikt_hlutfall))
    # tékkum fjölda UNK
    fjoldi_unk = len([x for x in min_lemmun if x[2] == "UNK"]) / len(min_lemmun)
    print("\nHlutfall ómarkaðra lesmálsorða (UNK):", prosenta(fjoldi_unk))
```

```
ver1()
# ver2a()
ver2c()
```

Úttak

```
    Lemmald, dæmi um tvær vitleysur:
[('en', 'c', 'en'), ('vantar', 'sfg3en', 'vanta'), ('enn', 'aa', 'enn'), ('undirtitilinn', 'nkeog', 'undirtitilu

[('.', '.', '.'), ('Útvarpsþættirnir', 'nkfng', 'útvarpsþáttur'), ('Áfangar', 'nkfn-s', 'áfangur'), (',', ',', '

Nefnir, dæmi um vitleysu:
[(',', ',', ',', ','), ('sem', 'ct', 'sem'), ('er', 'sfg3en', 'vera'), ('soldil', 'aa', 'soldil'), ('miðbæjarrotta',
Ólík lemmun:
Við sjáum að Nefnir breytir ekki kr. á meðan Lemmald lemmar kr. sem króna
------------
Lemmald: ('liðið', 'nheng', 'liði')
Nefnir: ('liðið', 'nheng', 'lið')
Lemmald: ('AC', 'e', 'ac')
Nefnir: ('AC', 'e', 'AC')
Lemmald: ('Milan', 'e', 'milan')
Nefnir: ('Milan', 'e', 'Milan')
Lemmald: ('um', 'aa', 'um')
Nefnir: ('um', 'aa', 'um')
Lemmald: ('4', 'ta', '4')
Nefnir: ('4', 'ta', '4')
Lemmald: ('milljarða', 'nkfo', 'milljarður')
Nefnir: ('milljarða', 'nkfo', 'milljarður')
Lemmald: ('ísl.', 'as', 'íslenskur')
Nefnir: ('ísl.', 'as', 'ísl.')
Lemmald: ('kr.', 'as', 'króna')
Nefnir: ('kr.', 'as', 'kr.')
Lemmald: ('Varaformaður', 'nken', 'varaformaður')
Nefnir: ('Varaformaður', 'nken', 'varaformaður')
Lemmald: ('félagsins', 'nheeg', 'félag')
Nefnir: ('félagsins', 'nheeg', 'félag')

Hlutföll:
---------
Hlutfall lesmálsorða sem eru lemmuð eins: 96.4%
Hlutfall lesmálsorða sem eru lemmuð ólíkt: 3.6%

Eins lemmun en hvort tveggja vitlaust:
Orðmyndin 'hægt' er eins lemmuð en við fáum ranga niðurstöðu þar sem orðið
er rangt markað. Í þessu samhengi er þetta atviksorð en er markað sem lýsingarorð.
-------------------------------------
Lemmald: ('.', '.', '.')
Nefnir: ('.', '.', '.')
Lemmald: ('Sama', 'fbkeo', 'samur')
Nefnir: ('Sama', 'fbkeo', 'sami')
Lemmald: ('dag', 'nkeo', 'dagur')
Nefnir: ('dag', 'nkeo', 'dagur')
Lemmald: ('og', 'c', 'og')
Nefnir: ('og', 'c', 'og')
Lemmald: ('leikir', 'nkfn', 'leikur')
Nefnir: ('leikir', 'nkfn', 'leikur')
Lemmald: ('fara', 'sfg3fn', 'fara')
Nefnir: ('fara', 'sfg3fn', 'fara')
Lemmald: ('fram', 'aa', 'fram')
Nefnir: ('fram', 'aa', 'fram')
```

```
Lemmald: ('er', 'sfg3en', 'vera')
Nefnir: ('er', 'sfg3en', 'vera')
Lemmald: ('hægt', 'lhensf', 'hægur')
Nefnir: ('hægt', 'lhensf', 'hægur')
Lemmald: ('að', 'cn', 'að')
Nefnir: ('að', 'cn', 'að')

Mismunandi lemmun og hvort tveggja vitlaust:
Orðmyndin 'fyrst' er rangt lemmuð í báðum tilvikum, ætti aö vera 'snemma'
--------------------------------------
Lemmald: ('heitið', 'nheng', 'heiti')
Nefnir: ('heitið', 'nheng', 'heiti')
Lemmald: ('kom', 'sfg3eþ', 'koma')
Nefnir: ('kom', 'sfg3eþ', 'koma')
Lemmald: ('fyrst', 'aae', 'fyrr')
Nefnir: ('fyrst', 'aae', 'snemmt')
Lemmald: ('fram', 'aa', 'fram')
Nefnir: ('fram', 'aa', 'fram')
Lemmald: ('í', 'aþ', 'í')
Nefnir: ('í', 'aþ', 'í')

Mín lemmun og nefnir:
Hlutfall lesmálsorða sem eru lemmuð eins: 67.6%
Hlutfall lesmálsorða sem eru lemmuð ólíkt: 32.4%

Mín lemmun og lemmald:
Hlutfall lesmálsorða sem eru lemmuð eins: 65.9%
Hlutfall lesmálsorða sem eru lemmuð ólíkt: 34.1%

Hlutfall ómarkaðra lesmálsorða (UNK): 28.7%
```