Projekt 3: Geodotazník

Obsah

Vypracovanie projektu Geodotazník

Predpokladaná časová náročnosť: 8hod
Skutočná časová náročnosť: 6hod

Úvod

Cieľom úlohy je spracovať odpovede učiteľov geografie na dve otvorené otázky v dotazníku. Z dát (celkovo 2 x 603 odpovedí) je potrebné získať najfrekventovanejšie odpovede a tie, ktoré by mohli byť zaujímavé pre ostatných učiteľov napriek tomu, že nie sú až tak časté. Presné znenie otázok a jednotlivé požadované výsledky vidíme v tabuľke 1. Na obrázku 1 zas vidíme ukážku dát v programe *Excel*. Na vypracovanie projektu použijeme programovací jazyk *Python*.

	Znenie otázky	Cieľ úlohy
Otázka 1	Které úlohy jsou při práci s mapami v atlase mezi vašimi žáky nejoblíbenější?	(1) Získať najčastejšie aktivity (2) Získať menej časté zaujímavé aktivity
Otázka 2	Které další zdroje map využíváte?	(1) Získať najfrekventovanejšie zdroje máp iné ako atlasy

Tabuľka 1: Presné znenie otázok a cieľ jednotlivých úloh.

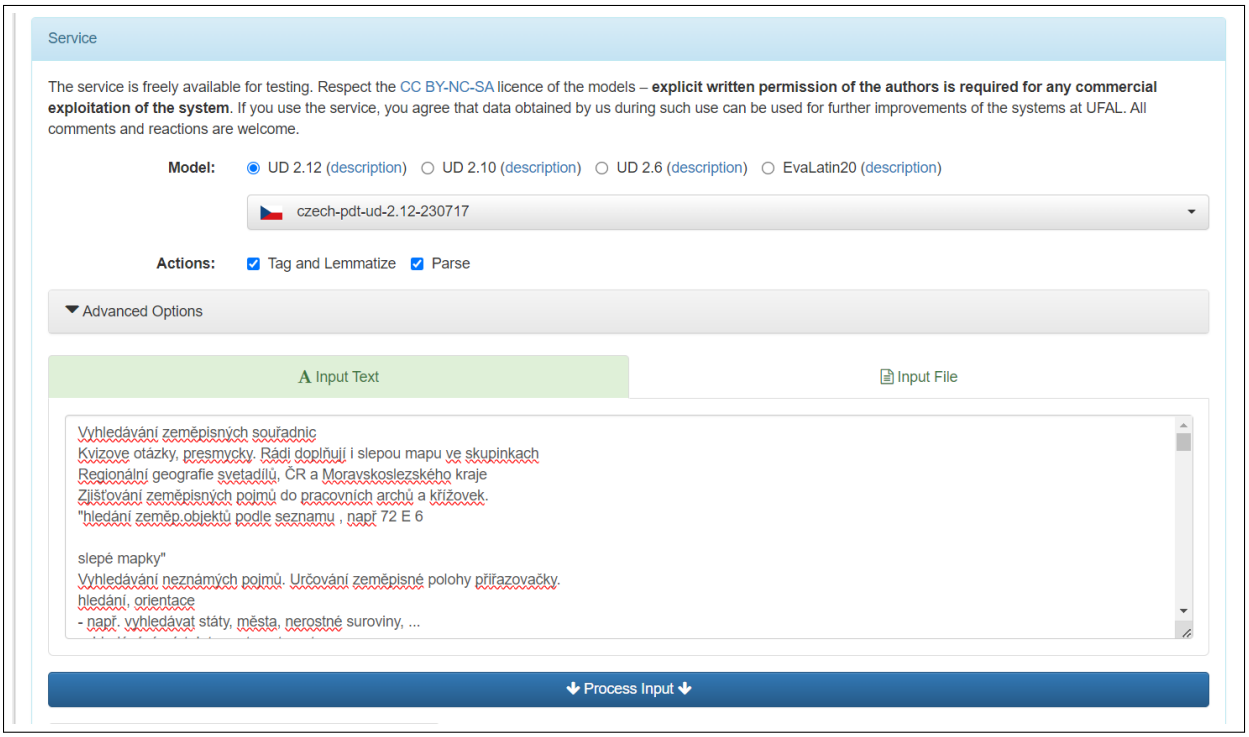
8_Které úlohy jsou při práci s mapami v atlase mezi vašimi žáky nejoblíbenější?	20_Které další zdroje map využíváte?
Vyhledávání zeměpisných souřadnic	
Kvizove otázky, presmycky. Rádi doplňují i slepou mapu ve skupinkach	Turistické mapy Mapu sídelMapy hustoty zalidnění
Regionální geografie svetadílů, ČR a Moravskoslezského kraje	Interaktivní mapy
Zjišťování zeměpisných pojmů do pracovních archů a křížovek.	Mapové aplikace
hledání zeměp.objektů podle seznamu , např 72 E 6slepé mapky	mapy.czGoogle Earth
Vyhledávání neznámých pojmů. Určování zeměpisné polohy přiřazovačky.	
hledání, orientace	Seznam - mapyGoogle Earth
- např. vyhledávat státy, města, nerostné suroviny, ...	webové stránky - např. soutěžě
vyhledávání míst, interpretace temat. map	google mapsgoogle earthmapy.cz
porovnávání údajů o státech a městech	mapové listy
Určování světových míst dle zeměpisných souřadnic...	
polohopis - práce s obrys.mapou	www.mapy.cz, googlemap
charakterizuj oblast nebo stát podle atlasu (co vše z atlasu můžete vyčíst...)vyhledávár	slepé mapyNárodní geoportál Inspire
Vyhledávání zeměpisných souřadnic ;).	Google maps

Obr. 1: Ukážka dát v programe *Excel*.

Metóda

Otázka 1: Obľúbené aktivity

V tejto časti sa pokúsime o analýzu obĺúbených aktivít. Avšak skôr než ju urobíme, musíme extrahovať jednotlivé aktivity a ich frekvencie. Keďže odpovede sú do veľkej miery heterogénne (nedodržiavajú diakritiku, slová sa vyskytujú v rôznych flektívnych tvaroch, atď.), potrebujeme ich nejak zhomogenizovať. Na to nám poslúži voľne dostupný online NLP nástroj *UDPipe* (`https://lindat.mff.cuni.cz/services/udpipe/`), ktorý dokáže tokenizovať, lematizovať, a určiť závislostnú syntax. Pokiaľ do pola pre vstup vložíme odpovede na otázku 1 (viď obrázok 2) a stlačíme tlačidlo *Process Input*, získame CoNLL-U súbor s potrebnými dátami, ktorý nám bude ďalej slúžiť pri extrakcii aktivít pomocou *Pythonu*.



Obr. 2: Vyplnené webové rozhranie aplikácie *UDPipe*.

Takto získaný CoNLL-U súbor spracujeme pomocou python kódu dostupného v prílohe 1. Spracované dáta potom nahráme do Excelu, a vytvoríme z nich tabuľky. Ako vidíme na obrázku 3, dostávame 4 tabuľky slov alebo ich kombinácií a ich frekvencií. Tabuľky obsahujú lemma tvary, bi, tri a quadgramy lemma tvarov. Tabuľky si pre prehľadnosť zoradíme podľa frekvencie od najväčšej po najmenšiu. Aj po takejto jednoduchej úprave už na prvý pohľad vidíme možné najoblúbenejšie aktivity.

V tabuľke 2 môžeme vidieť 50 najfrekventovanejších slov a slovných spojení v základnom tvare spolu s ich frekvenciami. Zdá sa, že najčastejšie aktivity sú: orientácia na mape, určovanie zemepisnej polohy, vyhľadávanie štátu, práca s atlasom, práca s indexom/registrom,

lemma	freq	lemma	freq	lemma	freq	lemma	freq
vyhledává	180	('práce', 's')	46	('orientace', 'na', 'mapa')	20	('orientace', 'na', 'mapa', 'vyhledávání')	5
mapa	164	('slepý', 'mapa')	37	('do', 'slepý', 'mapa')	14	('vyhledávání', 'stát', 'a', 'jeho')	3
a	123	('na', 'mapa')	27	('určování', 'zeměpisný', 'poloha'	12	('stát', 'a', 'jeho', 'hlavní')	3
na	99	('vyhledávání', 'informac	24	('práce', 's', 'atlas')	11	('a', 'jeho', 'hlavní', 'město')	3
v	91	('zeměpisný', 'poloha')	22	('práce', 's', 'rejstřík')	11	('zakreslování', 'do', 'slepý', 'mapa')	3
s	90	('orientace', 'na')	21	('mapa', 'vyhledávání', 'pojem')	6	('práce', 's', 'tematický', 'mapa')	3
být	76	('vyhledávání', 'pojem')	18	('na', 'mapa', 'vyhledávání')	5	('práce', 's', 'rejstřík', 'vyhledávání')	3
hledání	72	('mapa', 'vyhledávání')	17	('s', 'tematický', 'mapa')	5	('ten', 'se', 'muset', 'zeptat')	3
stát	71	('hlavní', 'město')	17	('místo', 'podle', 'souřadnice')	4	('práce', 's', 'tématický', 'mapa')	2
pojem	63	('s', 'atlas')	17	('vyhledávání', 'stát', 'a')	4	('při', 'práce', 's', 'atlas')	2
práce	57	('zeměpisný', 'souřadnic	16	('stát', 'a', 'jeho')	4	('mapa', 'vyhledávání', 'zadaný', 'pojem')	2
informace	55	('určování', 'zeměpisný')	15	('jeho', 'hlavní', 'město')	4	('rád', 'pracovat', 's', 'mapa')	2
místo	52	('do', 'slepý')	14	('orientace', 'v', 'mapa')	4	('pojem', 'do', 'slepý', 'mapa')	2
zeměpisný	51	('v', 'atlas')	14	('vyhledávání', 'zadaný', 'pojem')	4	('do', 'slepý', 'mapa', 'vyhledávání')	2
atlas	51	('vyhledávání', 'místo')	13	('slepý', 'mapa', 'vyhledávání')	4	('slepý', 'mapa', 'vyhledávání', 'pojem')	2
město	50	('časový', 'pásmo')	13	('práce', 's', 'mapa')	4	('žák', 'mít', 'například', 'obrysový')	2
se	48	('s', 'rejstřík')	12	('vyhledávání', 'zeměpisný', 'souř	3	('mít', 'například', 'obrysový', 'mapka')	2
ten	45	('být', 'ten')	11	('tvorba', 'vlastní', 'mapa')	3	('do', 'slepý', 'mapa', 'podle')	2
souřadnice	43	('s', 'mapa')	10	('vyhledávání', 'hlavní', 'město')	3	('slepý', 'mapa', 'podle', 'atlas')	2
do	39	('mapa', 'hledání')	9	('hlavní', 'město', 'stát')	3	('místo', 'orientace', 'na', 'mapa')	2
orientace	39	('v', 'mapa')	9	('a', 'jeho', 'hlavní')	3	('do', 'slepý', 'mapa', 'hra')	2
žák	39	('tematický', 'mapa')	9	('pracovat', 's', 'mapa')	3	('se', 'muset', 'zeptat', 'žák')	2
slepý	38	('stát', 'a')	8	('vyhledávání', 'informace', 'o')	3	('práce', 's', 'atlas', 'bavit')	2
který	36	('vlastní', 'mapa')	8	('zakreslování', 'do', 'slepý')	3	('na', 'mapa', 'vyhledávání', 'pojem')	2
podle	35	('z', 'mapa')	8	('práce', 's', 'tematický')	3	('vyhledávání', 'zadaný', 'pojem', 'vyhledáv	2
poloha	33	('místo', 'podle')	8	('odpověď', 'na', 'otázka')	3	('mapa', 'orientace', 'na', 'mapa')	2

Obr. 3: Ukážka získaných dát z CoNLL-U súboru pomocou *Python* kódu.

zakreslovanie do slepej mapy, vyhladávanie zadaného pojmu, určovanie miesta podľa jeho súradníc, určovanie zemepisných súradníc miesta, hľadanie – skupinová práca, určovanie hlavného mesta štátu, súťaženie žiakov, práca s politickou mapou, vlastná mapa, vyhľadávanie podľa registra, regionálna geografia, a iné.

V tabuľke 3 môžeme vidieť 50 najmenej frekventovaných slov a slovných spojení. Práve tu by sme mohli nájšť aktivity, ktoré síce nie sú bežné medzi ostatnými vyučujúcimi, no napriek tomu sa niektorým zdajú ako vhodné či zábavné pre žiakov. Sem teda patria činnosti ako: užšia regionálna geografia (Moravsko-sliezsky kraj), vyhladávanie zdrojov nerastných surovín, kvízové otázky. Okrem týchto sa v dotazníku vyskytujú ešte ďalšie aktivity, ako je napríklad vytváranie vlastného resumé či popis trasy.

	Lemma	*freq*	Bigram	*freq*	Trigram	*freq*	Quadgram	*freq*
1	vyhledávání	180	práce s	46	orientace na mapa	20	orientace na mapa vyhledávání	5
2	mapa	164	slepý mapa	37	do slepý mapa	14	vyhledávání stát a jeho	3
3	a	123	na mapa	27	určování zeměpisný poloha	12	stát a jeho hlavní	3
4	na	99	vyhledávání informace	24	práce s atlas	11	a jeho hlavní město	3
5	v	91	zeměpisný poloha	22	práce s rejstřík	11	zakreslování do slepý mapa	3
6	s	90	orientace na	21	mapa vyhledávání pojem	6	práce s tematický mapa	3
7	být	76	vyhledávání pojem	18	na mapa vyhledávání	5	práce s rejstřík vyhledávání	3
8	hledání	72	mapa vyhledávání	17	s tematický mapa	5	ten se muset zeptat	3
9	stát	71	hlavní město	17	místo podle souřadnice	4	práce s tématický mapa	2
10	pojem	63	s atlas	17	vyhledávání stát a	4	při práce s atlas	2
11	práce	57	zeměpisný souřadnice	16	stát a jeho	4	mapa vyhledávání zadaný pojem	2
12	informace	55	určování zeměpisný	15	jeho hlavní město	4	rád pracovat s mapa	2
13	místo	52	do slepý	14	orientace v mapa	4	pojem do slepý mapa	2
14	zeměpisný	51	v atlas	14	vyhledávání zadaný pojem	4	do slepý mapa vyhledávání	2
15	atlas	51	vyhledávání místo	13	slepý mapa vyhledávání	4	slepý mapa vyhledávání pojem	2
16	město	50	časový pásmo	13	práce s mapa	4	žák mít například obrysový	2
17	se	48	s rejstřík	12	vyhledávání zeměpisný souřadnice	3	mít například obrysový mapka	2
18	ten	45	být ten	11	tvorba vlastní mapa	3	do slepý mapa podle	2
19	souřadnice	43	s mapa	10	vyhledávání hlavní město	3	slepý mapa podle atlas	2
20	do	39	mapa hledání	9	hlavní město stát	3	místo orientace na mapa	2
21	orientace	39	v mapa	9	a jeho hlavní	3	do slepý mapa hra	2
22	žák	39	tematický mapa	9	pracovat s mapa	3	se muset zeptat žák	2
23	slepý	38	stát a	8	vyhledávání informace o	3	práce s atlas bavit	2
24	který	36	vlastní mapa	8	zakreslování do slepý	3	na mapa vyhledávání pojem	2
25	podle	35	z mapa	8	práce s tematický	3	vyhledávání zadaný pojem vyhledávání	2
26	poloha	33	místo podle	8	odpověď na otázka	3	mapa orientace na mapa	2
27	soutěž	32	informace o	8	určování zeměpisný souřadnice	3	země orientace na mapa	2
28	určování	30	hledání místo	7	s rejstřík vyhledávání	3	vyhledávání informace v tematický	2
29	například	29	podle souřadnice	7	ten se muset	3	informace v tematický mapa	2
30	z	29	vyhledávání stát	7	se muset zeptat	3	práce s obecně zeměpisný	2
31	najít	28	pracovní list	7	s atlas pracovat	3	s obecně zeměpisný mapa	2
32	on	26	na rychlost	7	s politický mapa	3	práce s politický mapa	2
33	úloha	25	souřadnice vyhledávání	7	zeměpisný souřadnice vyhledávání	3	určování zeměpisný poloha a	2
34	rejstřík	25	město stát	6	na malý jednička	3	místo podle zeměpisný souřadnice	2
35	mít	23	vyhledávání v	6	vyhledávání informace v	3	práce s slepý mapa	2
36	řeka	22	soutěž vyhledávání	6	doplňování slepý mapa	3	orientace na mapa hledání	2
37	křížovka	21	vyhledávání nový	6	město a stát	3	vyhledávání zeměpisný souřadnice kvizove	1
38	různý	21	orientace v	6	na fyzický mapa	3	zeměpisný souřadnice kvizove otázka	1
39	vědět	20	na základ	6	informace o stát	3	souřadnice kvizove otázka presmycek	1
40	o	18	odpověď na	6	o stát a	2	rád doplňovat i slepý	1
41	nebo	18	ten být	6	polohopis práce s	2	doplňovat i slepý mapa	1
42	hlavní	18	pojem a	6	slepý mapa hledání	2	i slepý mapa v	1
43	porovnávání	17	soutěž v	6	soutěž vyhledávání nový	2	slepý mapa v skupinka	1
44	co	17	mapa v	5	skupinový práce hledání	2	mapa v skupinka regionální	1
45	vyhledat	17	zeměpisný pojem	5	práce s tématický	2	v skupinka regionální geografie	1
46	pohoří	17	který se	5	s tématický mapa	2	skupinka regionální geografie svetadíl	1
47	daný	16	být se	5	soutěž vyhledávání v	2	regionální geografie svetadíl čr	1
48	zadaný	16	mapa a	5	při práce s	2	geografie svetadíl čr a	1
49	svět	16	tématický mapa	5	vyhledávání informace čtení	2	svetadíl čr a moravskoslezský	1
50	povrch	16	a jeho	5	čtení z mapa	2	čr a moravskoslezský kraj	1

Tabuľka 2: 50 najčastejších lemma tvarov, bigramov, trigramov a quadgramov spolu s ich frekvenciami.

	Lemma	*freq*	Bigram	*freq*	Trigram	*freq*	Quadgram	*freq*
1	kvizove	1	souřadnice kvizove	1	zeměpisný souřadnice kvizove	1	vyhledávání zeměpisný souřadnice kvizove	1
2	presmycek	1	kvizove otázka	1	souřadnice kvizove otázka	1	zeměpisný souřadnice kvizove otázka	1
3	skupinka	1	otázka presmycek	1	kvizove otázka presmycek	1	souřadnice kvizove otázka presmycek	1
4	svetadíl	1	rád doplňovat	1	rád doplňovat i	1	rád doplňovat i slepý	1
5	moravskoslezský	1	doplňovat i	1	doplňovat i slepý	1	doplňovat i slepý mapa	1
6	arch	1	i slepý	1	i slepý mapa	1	i slepý mapa v	1
7	zeměp	1	v skupinka	1	slepý mapa v	1	slepý mapa v skupinka	1
8	72	1	skupinka regionální	1	mapa v skupinka	1	mapa v skupinka regionální	1
9	e	1	geografie svetadíl	1	v skupinka regionální	1	v skupinka regionální geografie	1
10	přiřazovačka	1	svetadíl čr	1	skupinka regionální geografie	1	skupinka regionální geografie svetadíl	1
11	temato	1	čr a	1	regionální geografie svetadíl	1	regionální geografie svetadíl čr	1
12	obrys.mapa	1	a moravskoslezský	1	geografie svetadíl čr	1	geografie svetadíl čr a	1
13	charakteriovat	1	moravskoslezský kraj	1	svetadíl čr a	1	svetadíl čr a moravskoslezský	1
14	nísto	1	kraj zjišťování	1	čr a moravskoslezský	1	čr a moravskoslezský kraj	1
15	vyznačovat	1	zjišťování zeměpisný	1	a moravskoslezský kraj	1	a moravskoslezský kraj zjišťování	1
16	silný	1	pracovní arch	1	moravskoslezský kraj zjišťování	1	moravskoslezský kraj zjišťování zeměpisný	1
17	exkluzivita	1	arch a	1	kraj zjišťování zeměpisný	1	kraj zjišťování zeměpisný pojem	1
18	všimnout	1	a křížovka	1	zjišťování zeměpisný pojem	1	zjišťování zeměpisný pojem do	1
19	preference	1	hledání zeměp	1	zeměpisný pojem do	1	zeměpisný pojem do pracovní	1
20	vyžadovat	1	objekt podle	1	pojem do pracovní	1	pojem do pracovní arch	1
21	návrh	1	seznam například	1	do pracovní arch	1	do pracovní arch a	1
22	možný	1	například 72	1	pracovní arch a	1	pracovní arch a křížovka	1
23	výskyt	1	72 e	1	arch a křížovka	1	objekt podle seznam například	1
24	deigram	1	e 6	1	objekt podle seznam	1	podle seznam například 72	1
25	vytvíření	1	slepý mapka	1	podle seznam například	1	seznam například 72 e	1
26	resumé	1	mapka vyhledávání	1	seznam například 72	1	například 72 e 6	1
27	pozorování	1	poloha přiřazovačka	1	například 72 e	1	slepý mapka vyhledávání známý	1
28	poznatek	1	orientace například	1	72 e 6	1	mapka vyhledávání známý pojem	1
29	porovnávací	1	například vyhledávat	1	slepý mapka vyhledávání	1	určování zeměpisný poloha přiřazovačka	1
30	navazující	1	město nerostný	1	mapka vyhledávání známý	1	hledání orientace například vyhledávat	1
31	b	1	surovina vyhledávání	1	vyhledávání známý pojem	1	orientace například vyhledávat stát	1
32	c	1	místo interpretace	1	zeměpisný poloha přiřazovačka	1	například vyhledávat stát město	1
33	d	1	interpretace temato	1	hledání orientace například	1	vyhledávat stát město nerostný	1
34	destatz	1	mapa porovnávání	1	orientace například vyhledávat	1	stát město nerostný surovina	1
35	výchozí	1	porovnávání údaj	1	například vyhledávat stát	1	město nerostný surovina vyhledávání	1
36	cílový	1	údaj o	1	vyhledávat stát město	1	nerostný surovina vyhledávání místo	1
37	kvízový	1	a město	1	stát město nerostný	1	surovina vyhledávání místo interpretace	1
38	kvalitní	1	město určování	1	město nerostný surovina	1	vyhledávání místo interpretace temato	1
39	rozhodování	1	určování světový	1	nerostný surovina vyhledávání	1	mapa porovnávání údaj o	1
40	platnost	1	světový místo	1	surovina vyhledávání místo	1	porovnávání údaj o stát	1
41	tvrzení	1	dle zeměpisný	1	vyhledávání místo interpretace	1	údaj o stát a	1
42	správnost	1	souřadnice polohopis	1	místo interpretace temato	1	o stát a město	1
43	rozhodnout	1	s obrys.mapa	1	mapa porovnávání údaj	1	stát a město určování	1
44	několik	1	charakteriovat oblast	1	porovnávání údaj o	1	a město určování světový	1
45	nastejno	1	oblast nebo	1	údaj o stát	1	město určování světový místo	1
46	jednat	1	všechen z	1	stát a město	1	určování světový místo dle	1
47	určitě	1	atlas moci	1	a město určování	1	světový místo dle zeměpisný	1
48	pochopení	1	lokalita vyhledávání	1	město určování světový	1	místo dle zeměpisný souřadnice	1
49	souřadnicový	1	místo do	1	určování světový místo	1	dle zeměpisný souřadnice polohopis	1
50	systém	1	hledání nísto	1	světový místo dle	1	zeměpisný souřadnice polohopis práce	1

Tabuľka 3: 50 najmenej častých lemma tvarov, bigramov, trigramov a quadgramov spolu s ich frekvenciami.

Otázka 2: Iné zdroje máp

Cieľom úlohy je získať názvy iných zdrojov máp ako atlasy. Môže sa jednať o názvy serverov, internetových stránok či spoločností. Vďaka tomu, že hľadáme vlastné názvy je riešenie tejto úlohy variáciou na známy problém extrakcie informacií – *named entity recognition* (rozpoznanie menných entít).

Pre rýchle a jednoduché spracovanie menných entít si zavedieme definíciu mennej entity ako akéhokoľvek slova, ktoré začína veľkým písmenom. K menným entitám taktiež priradíme webstránky – slová, ktoré obsahujú **.** (symbol bodky). Táto definícia avšak nie je dokonalá a medzi menné entity zaradí aj tie, ktoré tam patriť nemajú, napr. začiatky viet, a vynechá aj niektoré tie, ktoré tam patriť majú (chybne zapísané názvy spoločností s malým písmenom). Keďže názvy spoločností či serverov môžu byť viacslovné, budeme skúmať aj n-gramy (2 až 4). Zameriame sa však len na tie, ktoré obsahujú aspoň jedno slovo s veľkým písmenom (našou mennou entitou).

Odpovede na druhú otázku si najprv uložíme do oddeleného *Excel* súboru, aby sme mohli s nimi jednoduchšie pracovať. Dáta spracujeme pomocou programovacieho jazyka *Python* a knižníc *string*, *pandas*, *collections*, *nltk*, *os* a *re*. Celý proces spracovania aj s komentármi môžeme vidieť v prílohe 2 v podobe *Jupyter Notebooku*. Spracované dáta si potom nahráme do *Excelu*, kde s nimi pre väčšiu prehľadnosť ďalej pracujeme. Postup ich spracovania v *Exceli* je totožný ako pri prvej úlohe – vytvoríme si tabuľku a zoradíme ju podľa frekvencií.

Za najfrekventovanejšie iné (ako atlas) zdroje máp môžeme považovať nasledovné názvy služieb a webstránok: Mapy Google, Google Earth, Mapy.cz, gisonline.cz, Wikipedia, Národní geoportál INSPIRE, Česká informační agentura životního prostředí (CENIA), Toporopa.eu, skolnimapy.cz, oldmaps.geolab.cz, worldmapper.org, umimefakta.cz, bergsteigen.com, geology.cz. Tieto a iné frekventované možné menné entity môžeme vidieť v tabuľke 4. Menej frekventované zdroje sú: cestovani.idnes.cz, worldometers.info, gapminder.org, thetruesize.com.

	1-slovný názov	*freq*	2-slovný názov	*freq*	3-slovný názov	*freq*	4-slovný názov	*freq*	webstránka	*freq*
1	google	86	google earth	37	google maps mapycz	9	google earth google mapy	5	mapy.cz	133
2	earth	31	google maps	22	mapycz google earth	8	mapycz google earth google	3	www.mapy	8
3	mapycz	30	mapycz google	17	google earth google	8	seznam mapy google earth	2	maps.google	6
4	mapy	19	google mapy	16	google earth mapycz	6	google earth webové stránky	2	maps.com	5
5	slepé	13	slepé mapy	10	earth google mapy	5	google maps mapycz google	2	google.cz	5
6	maps	11	earth google	8	mapycz google maps	5	maps mapycz google earth	2	slepemapy.cz	5
7	internet	9	mapy google	6	seznam mapy google	3	google earth google earth	2	www.umimefakta	4
8	používám	8	earth mapycz	6	mapy google earth	3	earth google mapy mapy	2	umimefakta.cz	4
9	seznam	7	maps mapycz	6	google maps google	3	google mapy mapy na	2	google.maps	3
10	online	5	seznam mapy	5	mapycz google mapy	3	seznam mapy google mapy	2	www.google	3
11	nástěnné	5	mapy mapycz	4	slepé mapy z	3	mapycz google maps google	2	zemepis.com	3
12	mapové	4	maps google	4	google earth webové	2	slepé mapy z internetu	2	www.zemepis	2
13	gis	4	mapy z	4	earth webové stránky	2	turistické mapy mapu sídel	1	google.com	2
14	gps	4	nástěnné mapy	4	google maps mapy	2	mapy mapu sídel mapy	1	otkymty5nw.nzmxndcwnq	2
15	pc	3	online mapy	3	maps mapycz google	2	mapu sídel mapy hustoty	1	thetruesize.com	2
16	wikipedie	3	mapy používám	3	earth google earth	2	sídel mapy hustoty zalidnění	1	geology.cz	2
17	využíváme	3	google mapycz	3	google mapy mapy	2	mapy hustoty zalidnění interaktivní	1	nti2nda1mq.nzg2mzqymq	2
18	internetové	3	google map	3	mapy google mapy	2	hustoty zalidnění interaktivní mapy	1	googlemaps.com	2
19	nejčastěji	3	mapy mapy	3	google mapy mapycz	2	zalidnění interaktivní mapy mapové	1	prb.org	2
20	nebo	3	mapy na	3	mapy na internetu	2	interaktivní mapy mapové aplikace	1	www.maps	1
21	streetview	3	mapy slepé	3	internetu výukové programy	2	mapy mapové aplikace mapycz	1	mapa.google	1
22	world	3	turistické mapy	2	google mapy internet	2	mapové aplikace mapycz google	1	mtawmjqwnzu.mjuwmjm1mtc	1
23	např	3	earth webové	2	nástěnné mapy mapy	2	aplikace mapycz google earth	1	1.54	1
24	web	3	na pc	2	mapy nástěnné mapy	2	mapycz google earth seznam	1	1.43	1
25	čhmú	3	používám je	2	google maps pro	2	google earth seznam mapy	1	mtuwndqyotc.ntgyndcyng	1
26	využívám	3	mapové portály	2	slepé mapy mapycz	2	earth seznam mapy google	1	online.seterra	1
27	různé	3	internet mapy	2	turistické mapy mapu	1	mapy google earth webové	1	en.wikipedia	1
28	spíše	3	mapy online	2	mapy mapu sídel	1	earth webové stránky např	1	www.slepemapy	1
29	když	3	nová škola	2	mapu sídel mapy	1	googlemap slepé mapy národní	1	6.tř	1
30	turistické	2	zdroje google	2	sídel mapy hustoty	1	slepé mapy národní geoportál	1	www.worldpopdata	1
31	mapu	2	google slepé	2	mapy hustoty zalidnění	1	mapy národní geoportál inspire	1	9.tř	1
32	interaktivní	2	mapycz mapy	2	hustoty zalidnění interaktivní	1	národní geoportál inspire google	1	europe.arounder	1
33	inspire	2	míst google	2	zalidnění interaktivní mapy	1	geoportál inspire google maps	1	worldatlas.com	1
34	země	2	podrobnější mapy	2	interaktivní mapy mapové	1	inspire google maps mapy	1	size.com	1
35	dále	2	používám často	2	mapy mapové aplikace	1	google maps mapy z	1	www.googlemapy	1
36	digitální	2	internetu výukové	2	mapové aplikace mapycz	1	k jednotlivým světadílům mapycz	1	bergsteigen.com	1
37	seterra	2	výukové programy	2	aplikace mapycz google	1	jednotlivým světadílům mapycz interaktivní	1	worldmapper.org	1
38	nová	2	stahuji si	2	google earth seznam	1	světadílům mapycz interaktivní úlohy	1	oldmaps.geolab	1
39	při	2	mapy internet	2	earth seznam mapy	1	mapycz interaktivní úlohy kdy	1	kontaminace.cenia	1
40	ezilon	2	na zemi	2	slepé mapy národní	1	interaktivní úlohy kdy žáci	1	geoportal.cenia	1
41	měření	2	mapy nástěnné	2	mapy národní geoportál	1	digitální interaktivní mapy mapycz	1	games.com	1
42	googleearth	2	když je	2	národní geoportál inspire	1	interaktivní mapy mapycz k	1	skolniatlassveta.cz	1
43	čúzk	2	mapy mapu	1	geoportál inspire google	1	mapy mapycz k nácviku	1	www.skolniatlassveta	1
44	wikipedii	2	mapu sídel	1	inspire google maps	1	mapycz k nácviku používání	1	sedac.ciesin	1
45	nj	2	sídel mapy	1	jednotlivým světadílům mapycz	1	plánování trasy na pc	1	cenia.cz	1
46	podrobnější	2	mapy hustoty	1	světadílům mapycz interaktivní	1	trasy na pc vytištěné	1	bbc.co	1
47	výukové	2	zalidnění interaktivní	1	mapycz interaktivní úlohy	1	na pc vytištěné mapy	1	free.com	1
48	stahuji	2	interaktivní mapy	1	interaktivní úlohy kdy	1	pc vytištěné mapy pak	1	google.earth	1
49	tištěné	2	mapy mapové	1	interaktivní mapy mapycz	1	legendy mapy z wikipedie	1	www.toporopa	1
50	zemi	2	mapové aplikace	1	mapy mapycz k	1	mapy z wikipedie pro	1	toporopa.eu	1

Tabuľka 4: 50 najfrekventovanejších možných názvov zdrojov (servery, stránky, spoločnosti).

Záver

Zadaná úloha je úplne riešiteľná. Pomocou *UDPipe* sa nám podarilo homogenizovať pôvodne nesúrodé dáta. Tento krok bol potrebný pre jednoduché riešenie prvej otázky. Bez tohto kroku by naša analýza bola zložitejšia, pretože odpovede obsahovali názvy jednotlivých aktivít v rôznych tvaroch a bez diakritiky. Frekvencie jednotlivých odpovedí sme získali pomocou programovacieho jazyka *Python* (kód, ktorý sme použili sa nachádza v prílohe 1 a prílohe 2 aj s vysvetlujúcimi komentármi). Výsledky sme spracovávali v *Exceli*.

Nakoľko bola naša metóda zameraná skôr na rýchlosť riešenia a počet výsledkov než na úplnú spolahlivosť a dôkladnosť, je možné, že sme nejaké aktivity a názvy zdrojov mohli prehliadnuť alebo vynechať. Domnievame sa však, že tento postup prináša dostatočne jednoducho a rýchlo veľký počet dobrých, reprezentatívnych výsledkov, ktoré sa dajú ďalej spracovávať. Precíznejšia metóda by si vyžadovala lepšie predspracovanie dát (napr. lepší lematizér) či vhodnejšie definovanie toho, aký tvar môžu mať názvy zdrojov.

Príloha 1: Jupyter Notebook Extrakcia aktivít

Extrakcia aktivít z odpovedí Geodotazníku

Import knižníc

[1]:
```
import string
import pandas as pd
from conllu import parse
from collections import Counter
from nltk import ngrams
from os import path
from itertools import chain
```

Nastavenie potrebných premenných pre ukladanie

[2]:
```
files_directory_path = r"CESTA_K_ZLOZKE_S_CONLLU_SUBOROM"
working_dir = r"CESTA_K_PRACOVNEJ_ZLOZKE"
conllu_file = r"processed.conllu"
tsv_output_file = r"tsv_output_file.tsv"
conllu_file_path = path.join(files_directory_path, conllu_file)
tsv_output_path = path.join(working_dir, tsv_output_file)
```

Nahratie conllu súboru - výsledok spracovania UDPipe

[3]:
```
with open(conllu_file_path, 'r', encoding="UTF-8") as input_file:
    data_string = input_file.read()
data = parse(data_string)
```

Extrakcia dát z Conllu súboru

[4]:
```
lemmas = [[token["lemma"] for token in data_row if token["lemma"] not in string.
 ↪punctuation] for data_row in data]  # získanie všetkých lemma tvarov bez interpunkcie
bigrams = [list(ngrams(lemma, 2)) for lemma in lemmas]  # vytvorenie n-gramov z␣
 ↪extrahovaných lemma tvarov bez interpunkcie
trigrams = [list(ngrams(lemma, 3)) for lemma in lemmas]
quadgrams = [list(ngrams(lemma, 4)) for lemma in lemmas]
```

Získanie frekvencií lemma tvarov a n-gramov

[5]:
```
lemmas_freqs = Counter(list(chain(*lemmas)))
bigrams_freqs = Counter(list(chain(*bigrams)))
trigrams_freqs = Counter(list(chain(*trigrams)))
quadgrams_freqs = Counter(list(chain(*quadgrams)))
```

[6]:
```
lemmas_df = pd.DataFrame(list(lemmas_freqs.items()), columns=["lemma", "freq"])
bigrams_df = pd.DataFrame(list(bigrams_freqs.items()), columns=["lemma", "freq"])
trigrams_df = pd.DataFrame(list(trigrams_freqs.items()), columns=["lemma", "freq"])
quadgrams_df = pd.DataFrame(list(quadgrams_freqs.items()), columns=["lemma", "freq"])
```

[10]:
```
trigrams_df.sort_values(by="freq", ascending=False).head() # ukazka dat
```

[10]:

	lemma	freq
134	(orientace, na, mapa)	20
73	(do, slepý, mapa)	14
32	(určování, zeměpisný, poloha)	12
210	(práce, s, atlas)	11

```
668           (práce, s, rejstřík)    11
```

Uloženie dát do tsv súboru

Dáta ukladáme postupne zámenou argumentu funkcie save_data_to_tsv.

[8]:
```
save_data_to_tsv = lambda pandas_dataframe: pandas_dataframe.
↪to_csv(tsv_output_path,sep="\t", index=False)
```

[9]:
```
save_data_to_tsv(quadgrams_df) # uložené dáta si prekopírovaním zobrazíme v Exceli
```

Príloha 2: Jupyter Notebook Extrakcia menných entít a webstránok

Extrakcia menných entít z odpovedí Geodotazníku

Import knižníc

[ ]:
```
import string
import pandas as pd
from collections import Counter
from nltk import ngrams
from os import path
import re
```

Nastavenie potrebných premenných pre ukladanie

[2]:
```
working_dir = r"CESTA_K_PRACOVNEJ_ZLOZKE"
file_name = r"NAZOV_SUBORU_S_DATAMI_Z_OTAZKY2_VO_WORKING_DIR.xlsx"
output_file_name = "ner_answers.tsv"
file_path = path.join(working_dir, file_name)   # cesta k excel súboru s odpoveďami na otázku 2
output_file_path = path.join(working_dir, output_file_name) # cesta k súboru, kam budeme výsledky ukladať
```

[3]:
```
data = pd.read_excel(file_path,header=0) # načítanie dát
```

Vyčistenie dát

[4]:
```
def clear_data(row):
    """
    funkcia zmaže prebytočné znaky v odpovediach
    """
    row_str = row[0]
    row_str = row_str.replace("\n", " ")
    row_str = row_str.replace("\t", " ")
    row_str = row_str.replace("\xa0", " ")
    return row_str
```

[5]:
```
data = data.apply(clear_data, axis=1)
data.head() # ukážka dát
```

[5]:
```
0    Turistické mapy   Mapu sídel  Mapy hustoty zal...
1                                    Interaktivní mapy
2                                      Mapové aplikace
3                                 mapy.cz  Google Earth
4                           Seznam - mapy  Google Earth
dtype: object
```

Extrakcia možných menných entít

Možná menná entita – čokoľvek, čo začína veľkým písmenom

[6]:
```
def lower_case_words_in_list(lst):
    """
    funkcia zmení všetky slová v zozname na malé písmená
    """
    return [word.lower() for word in lst]
```

```
def extract_possible_named_entities_ngrams(data_str, n=1,lw=False):
    """
    funkcia nájde všetky n-gramy, ktoré obsahujú možné menné entity (slová s veľkým␣
 ↪písmenom na začiatku)
    """
    possible_named_entities_ngrams = []
    for gram in ngrams(data_str.split(),n):
        ngram_string = " ".join(gram)
        word_with_upper_case_beginning =  r"[A-Z].+\s*"
        possible_named_entities_ngram = re.findall(string=ngram_string,
                                                   ␣
 ↪pattern=word_with_upper_case_beginning)
        if possible_named_entities_ngram:
            possible_named_entities_ngrams.append(ngram_string)
    if lw:
        possible_named_entities_ngrams =␣
 ↪lower_case_words_in_list(possible_named_entities_ngrams)
    return possible_named_entities_ngrams
```

[ ]:
```
data_str = " ".join(data.to_list()) # string so vsetkymi odpovedami
data_str_without_punctuation = data_str.translate(str.maketrans("","", string.
 ↪punctuation))
extract_poss_ner_ngram = lambda n:␣
 ↪extract_possible_named_entities_ngrams(data_str_without_punctuation,
                                                                    n,
                                                                    lw=True)
```

[7]:
```
poss_named_entities = pd.DataFrame(Counter(extract_poss_ner_ngram(1)).items())
poss_named_entities_bigrams = pd.DataFrame(Counter(extract_poss_ner_ngram(2)).items())
poss_named_entities_trigrams =  pd.DataFrame(Counter(extract_poss_ner_ngram(3)).
 ↪items())
poss_named_entities_quadgrams = pd.DataFrame(Counter(extract_poss_ner_ngram(4)).
 ↪items())
```

Extrakcia webových stŕanok

Webová stŕanka – čokoľvek, čo v sebe zahŕňa symbol . (bodka)

[8]:
```
def find_web_pages(data_str):
    """Nájde všetky výskyty slov, ktoré obsahujú bodku"""
    return re.findall(string=data_str, pattern=r"\w+?\.\w+")
```

[9]:
```
web_pages = pd.DataFrame(Counter(lower_case_words_in_list(find_web_pages(data_str))).
 ↪items())
```

[10]:
```
web_pages.sort_values(by=1, ascending=False).head() # ukážka dát
```

[10]:

	0	1
0	mapy.cz	133
1	www.mapy	8
13	maps.google	6
2	maps.com	5
7	google.cz	5

Uloženie dát do tsv súboru

Dáta ukladáme postupne zámenou argumentu funkcie save_data_to_tsv.

[11]:
```
save_data_to_tsv = lambda pandas_dataframe: pandas_dataframe.
↪to_csv(output_file_path,sep="\t", index=False, header=False)
```

[12]:
```
save_data_to_tsv(web_pages) # uložené dáta si prekopírovaním zobrazíme v Exceli
```