Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Pokročilé přístupy k DZD

Zadání semestrální práce

Zpráva o specifikaci podoby umělých dat

Ludmila Svobodová

2. dubna 2015

1 Popis domény a motivace generování dat

Jako doménu semestrální práce jsem zvolila čerpání nemocenských v České republice. Přínos ve vygenerování dat z této domény vidím v tom, že i když se dá o nemocenských a obecně nemocnosti v ČR najít hodně znalostí, jsou tyto znalosti roztroušené a ne vždy dané do bližších souvislostí. Uměle vygenerovaná data by tedy mohla odrážet doménové znalosti z širšího hlediska a zároveň by se do nich již daly zakomponovat i pokročilejší a ne na první pohled jasné vztahy v této doméně.

Doména je zajímavá i z pohledu její aktuálnosti, protože zdraví obecně je poměrně dosti sledované a debatované téma. K dispozici je tedy velké množství aktuálních dat a statistik, ze kterých se dají znalosti o této doméně naučit.

Základní data o počtu nemocenských v ČR jsem se rozhodla zkoumat ve spojitosti s dalšími údaji, jako jsou vybrané skupiny diagnóz, věková struktura obyvatel, hrubé měsíční mzdy, nezaměstnanost a znečištění ovzduší. Mezi základní doménové znalosti tedy patří:

Vztahy pro 4FT-miner
• každý zdravotně pojištěný člověk byl v průměru 3,6 % roku na nemocenské, déle na nemocenské bývají ženy než muži [7] • nejdéle trvají pracovní neschopnosti s diagnózou zhoubných novotvarů a tuberkulózy, nejkratší dobu trvají nemocenské z důvodu nemoci dýchacích cest [10] • nemocenská z důvodu těhotenství, porodu a šestinedělí se vyskytuje převážně u žen do 45 let, nejpravděpodobnější je u žen od 25 do 35 let [12]

Vztahy pro CF-miner
• z celkového počtu nemocenských je absolutní většina z důvodu nemoci (87,7 % v roce 2013), zbývající případy připadají na pracovní a ostatní úrazy [5]

Vztahy pro KL-miner
• čím starší člověk je, tím je delší jeho doba strávená na nemocenské [12]

- největší počet pracovních neschopností mají muži od 20 do 39 let a ženy od 25 do 39 let [12]

- nejčastěji nemocní bývají lidé do 24 let [12]

2 Popis navržené struktury dat a rozdělení hodnot

2.1 Hlavní data

Skupina	Sloupec	Datový typ	Popis
Osoba	Věk	Integer	věk osoby
Osoba	Pohlaví	Text	pohlaví osoby
Osoba	Okres	Text	okres bydliště
Zaměstnání	Plat	Integer	plat osoby
Zaměstnání	Obor	Text	obor ekonomické činnosti
Zaměstnání	Okres	Text	územně-správní jednotka místa zaměstnání
Zaměstnání	Kraj	Text	územní jednotka místa zaměstnání
Nemocenská	Výskyt	Boolean	zda osoba za rok na nemocenské byla nebo ne
Nemocenská	Délka	Integer	doba trvání nemocenské
Nemocenská	Důvod	Text	důvod nemocenské nemoc/úraz
Nemocenská	Diagnóza	Text	zařazení nemocenské dle diagnózy

2.2 Data z externích zdrojů

Skupina	Podskupina	Sloupec	Datový typ
Okres	Zaměstnanost	Počet zaměstnanců	Integer
Okres	Dojíždění	Dojíždějící zaměstnanci	Integer
Okres	Nezaměstnanost	Nezaměstnanost	Float
Okres	Emise	Tuhé	Float
Okres	Emise	Oxid siřičitý	Float
Okres	Emise	Oxidy dusíku	Float
Okres	Emise	Oxid uhelnatý	Float

Sloupec	Popis	Zdroj
Počet zaměstnanců	počet zaměstnanců v okrese	[2]
Dojíždějící zaměstnanci	počet osob dojíždějících do okresu za prací	[10]
Nezaměstnanost	míra nezaměstnanosti	[5]
Tuhé	tuhé emise v ovzduší	[7]
Oxid siřičitý	množství Oxidu siřičitého obsaženého v ovzduší	[7]
Oxidy dusíku	množství Oxidů dusíku obsaženého v ovzduší	[7]
Oxid uhelnatý	množství Oxidu uhelnatého obsaženého v ovzduší	[7]

3 Popis navržených vzorů v datech

Umělá data by měla zohledňovat základní doménové znalosti. Doménové znalosti odpovídají obecnějším vztahům v doméně, avšak pro vygenerování konkrétních dat budou použity nejnovější statistiky, tedy převážně statistiky za rok 2013. Podle těchto statistik budou stanoveny konkrétní meze vztahů, které by umělá data v základu měla splňovat.

Cílem tedy je, aby umělá data splňovala doménové znalosti uvedené výše a tyto přidané vztahy:

Vztahy pro 4FT-miner
• nejvíce případů nemocenských s diagnózou nemoci dýchací soustavy je u řemeslníků a úředníků [12] • nejméně případů nemocenských s diagnózou nemoci dýchací soustavy je u řídících pracovníků a u kvalifikovaných pracovníků v zemědělství, lesnictví a rybářství [12] • nejvíce případů nemocenských s diagnózou těhotenství, porod a šestinedělí je u úřednic a pracujících v oblasti služeb a prodeje [12] • nejméně případů nemocenských s diagnózou těhotenství, porod a šestinedělí je u obsluhy strojů [12] • nejvíce případů nemocenských s diagnózou poranění, otravy aj. následky vnějších příčin je u řemeslníků a opravářů [12] • nejméně případů nemocenských s diagnózou poranění, otravy aj. následky vnějších příčin je u technických a odborných pracovníků [12] • nejvíce případů nemocenských s diagnózou novotvarů je u úředníků [12] • nejméně případů nemocenských s diagnózou novotvarů je u obsluh strojů [12]

Vztahy pro CF-miner
• nejdelší doba průměrné nemocenské je v okresu Šumperk a nejkratší je v Mladé Boleslavi a Praze [5] • nejvíce nemocenských je v okresech Prachatice a Šumperk [8] • nejméně nemocenských je v okresech Jeseník a Praha [8]

- nejdelší trvání nemocenské mají řídící pracovníci [12]
- nejkratší trvání nemocenské mají techničtí a odborní pracovníci [12]
- více případů nemocenské je u zaměstnanců s nižšími platy [11]
- čím starší člověk je, tím je delší jeho doba strávená na nemocenské [12]
- 71 % lidí je nemocných jednou, 20 % dvakrát a 6 % třikrát, 3 % lidí jsou nemocná více než třikrát [12]

- v Praze je nejméně nemocenských pro všechny typy diagnóz [12]
- nejvíce nemocenských z důvodu nemoci dýchacích cest je v Plzeňském kraji [12]
- nejvíce nemocenských z důvodu těhotenství, porodu a šestinedělí je v Olomouckém kraji [12]
- nejvíce nemocenských z důvodu novotvarů je v Karlovarském kraji [12]
- nejvíce nemocenských z důvodu poranění, otravy aj. následků vnějších příčin je v Jihočeském kraji [12]

Při generování umělých dat bude potřeba zajistit existenci výše uvedených vztahů, také se ale bude muset zajistit neexistence stejných nebo podobných vztahů u skupin, kde tyto vztahy být nemají.

Reference

[1] Český statistický úřad: Dojíždějící do zaměstnání a školy podle frekvence dojížďky, podle kraje a okresu dojížďky a podle pohlaví. [online] [cit. 2015-03-22]. Dostupné z: `http://tinyurl.com/o3va747`

[2] Český statistický úřad: Ekonomické subjekty podle počtu zaměstnanců. [online] [cit. 2015-03-22]. Dostupné z: `http://vdb.czso.cz/vdbvo/tabparam.jsp?cislotab=ORG9010UC&kapitola_id=6&voa=tabulka&go_zobraz=1&childsel0=3`

[3] Český statistický úřad: Emise základních znečišťujících látek. [online] [cit. 2015-03-22]. Dostupné z: `http://vdb.czso.cz/vdbvo/tabparam.jsp?voa=tabulka&cislotab=ZPR5012PU_OK&vo=null`

[4] Český statistický úřad: Míra registrované nezaměstnanosti, uchazeči s nárokem na podporu v nezaměstnanosti, uchazeči v rekvalifikaci. [online] [cit. 2015-03-22]. Dostupné z: `http://vdb.czso.cz/vdbvo/tabparam.jsp?voa=tabulka&cislotab=PRA5042PU_OK&vo=null`

[5] Český statistický úřad: Nemocensky pojištění a pracovní neschopnost. [online] [cit. 2015-03-22]. Dostupné z: `http://vdb.czso.cz/vdbvo/tabparam.jsp?voa=tabulka&cislotab=ZDR5022PU_OK&vo=null`

[6] Český statistický úřad: Obyvatelstvo ekonomicky aktivní podle věku, vzdělání, pohlaví a ekonomické aktivity. [online] [cit. 2015-03-22]. Dostupné z: `http://tinyurl.com/pe9yak7`

[7] Český statistický úřad: Průměrné procento pracovní neschopnosti. [online] [cit. 2015-03-22]. Dostupné z: http://vdb.czso.cz/vdbvo/tabparam.jsp?voa=tabulka&cislotab=ZDR5032PU_OK&vo=null

[8] Český statistický úřad: Věkové složení obyvatelstva. [online] [cit. 2015-03-22]. Dostupné z: http://vdb.czso.cz/vdbvo/tabparam.jsp?voa=tabulka&cislotab=OBY5062PU_OK&vo=null

[9] Český statistický úřad: Zaměstnanci a jejich průměrné hrubé měsíční mzdy podle CZ-NACE. [online] [cit. 2015-03-22]. Dostupné z: http://vdb.czso.cz/vdbvo/tabparam.jsp?vo=null&cislotab=PRA5091PU_KR&voa=tabulka&go_zobraz=1&childsel0=2&cas_2_21=2012

[10] J. Pirochová: Muskuloskeletární onemocnění z pohledu ČSSZ. [online] [cit. 2015-03-22]. Dostupné z: http://www.iheta.org/ext/files/47/Fit-for-Work-iHETA-Pirochova.pdf

[11] MPSV – odbor sociálního pojištění: Analýza vývoje nemocenského pojištění 2013. [online] [cit. 2015-03-22]. Dostupné z: http://www.mpsv.cz/files/clanky/12643/Analyza_2013.pdf

[12] Ústav zdravotnických informací a statistiky ČR: Ukončené případy pracovní neschopnosti pro nemoc a úraz 2012. [online] [cit. 2015-03-22]. Dostupné z: http://www.uzis.cz/system/files/uppn2012.pdf