

## SPARK

### Question 1

"Catholic, : 1 "I : 1 "John" : 1 "Think : 1 "adoptive : 1 "aggressive, : 1 "alcoholic : 1 "doctor : 1 "he : 1 "it : 1 "my : 1 "one : 1 "repo : 1 "threatened : 1 "was : 2 "were : 1 "without : 1 (/dʒɒbz/; : 1 (1922–1993), : 1 (1924–1986). : 1 (Arabic : 1 (CEO), : 1 (GUI). : 1 (b. : 2 ... : 3 1, : 1 1,000%."[12] : 1 15, : 1 17-year-old : 1 1930s : 1 1931), : 1 1932), : 1 1946.[12] : 1 1952.[12] : 1 1954 : 1 1954.[15] : 1 1955 : 1 1955, : 1 1955.[12] : 1 1960s.[4] : 1 1970s : 1 1972 : 1 1974 : 1 1976 : 1 1979, : 1 1980s. : 1 1983, : 1 1984. : 1 1985 : 1 1985.[9] : 1 1986.[10] : 1 1997 : 1 1997, : 1 2001, : 1 2003 : 1 2011) : 1 2011, : 1 23 : 1 24, : 2 5, : 2 Abdulfattah : 2 According : 1 After : 3 Although : 1 Alto, : 1 American : 2 And : 2 App : 1 Apple : 11 Apple, : 1 Arab : 1 Arabs : 1 Area : 2 Armenian : 1 As : 2 August : 1 Background : 1 Bay : 2 Beginning : 1 Beirut, : 1 Brennan, : 1 Buddhism.[6] : 1 But : 1 CEO : 3 Calvinist : 1 Carole : 1 Catholic : 1 Chrisann : 2 Clara : 8 Clara, : 1 Clara.[19] : 1 Coast : 2 College : 1 Company's : 1 Dean; : 1 Disney : 1 District : 1 Even : 1 FBI : 1 Family : 1 February : 2 Following : 1 Francisco : 6 Francisco's : 1 Francisco.[15] : 1 GUI, : 1 George : 1 German : 1 Germantown, : 1 Guard : 2 Hagopian : 1 Hagopian. : 1 He : 5 His : 2 Homs, : 2 I : 6 II, : 2 In : 6 Inc., : 1 India : 1 Isaacson, : 1 Ive : 1 James : 1 Jandali : 8 Jandali, : 1 Jandali : 1 Joanne : 5 Jobs : 23 Jobs's : 8 Jobs, : 1 Jonathan : 1 LSD : 2 LaserWriter, : 1 Laurene : 1 Lebanon, : 1 Lisa : 1 Lucas's : 1 Lucasfilm : 1 Mac : 2 Macintosh : 2 Many : 1 March : 1 Meanwhile, : 1 Michigan : 1 Middle-Eastern : 1 Midwest : 1 Mona : 1 Mr. : 1 Mrs. : 1 Muslim : 1 Muslim. : 1 NeXT's : 1 NeXT, : 1 NeXT. : 2 NeXTSTEP : 1 OS : 3 October : 2 PARC, : 1 Paul : 11 PhD : 1 Pixar, : 1 Pixar; : 1 Pixar; [3] : 1 Reed : 1 Reinhold : 1 San : 8 Schieble : 8 Schieble's : 2 Schieble. : 1 She : 1 Simpson, : 1 So : 1 States : 2 Steve : 6 Steve, : 1 Steven : 1 Store, : 3 Story—an : 1 Sunset : 1 Swiss : 1 Syria : 1 Syria. : 2 That's : 1 The : 6 They : 1 This : 1 Toy : 1 United : 2 University : 2 Unix-based : 1 Walt : 1 Walter : 1 War : 1 When : 2 While : 1 Wisconsin, : 1 Wisconsin. : 1 Wisconsin.[12][11][13] : 1 Wisconsin.[12][15] : 1 With : 1 Within : 1 World : 1 Wozniak : 2 Wozniak's : 1 X, : 1 Xerox : 1 Zen : 1 [Steve] : 1 a : 45 abortions : 1 abusive" : 1 acquaintance : 1 acquisition : 1 activist : 1 activities.[11] : 1 addition : 2 addition, : 2 additionally : 1 admitted : 1 adopt : 1 adoption : 4 adoption."[17] : 1 adoptions."[12] : 1 adoptive : 3 advertising : 1 after : 4 against : 1 age.[14] : 1 aggravate : 1 already : 1 also : 2 although : 1 am : 1 an : 10 and : 53 anyone : 1 are : 2 are, : 1 around : 1 arranged : 1 arrest : 1 as : 13 assistant : 1 at : 6 attempt : 1 attempts : 1 attend : 1 attended : 1 aware : 1 away : 1 babies, : 1 baby : 5 bank : 1 bank. : 1 bankruptcy. : 1 based : 1 bearing : 1 became : 2 because : 1 become : 1 been : 3 before : 1 before, : 1 being : 1 best : 1 bet : 1 biographer : 1 biographer, : 1 biological : 5 birth : 1 birth; : 1 blessed : 1 blind : 1 blue : 1 board : 1 bore : 1 born : 2 both : 2 boy : 1 breakthrough : 1 bring : 1 business : 2 but : 3 by : 4 campaign, : 1 candidate, : 1 care : 1 career : 1 cars, : 1 case, : 1 chairman, : 2 changed : 1 chief : 1 child : 2 chose : 1 closed : 1 closely : 1 co-founded : 1 co-founder : 2 collar : 1 college : 2 college.[12] : 1 college.[7] : 1 comment : 1 commercial : 1 company : 1 company, : 1 company; : 1 completely : 1 completely" : 1 computer : 3 computer-animated : 1 computer. : 1 computers : 1 computers. : 1 consented : 1 consider : 1 continued : 1 couple : 3 course : 1 court : 1 cultural : 1 cut : 1 dad. : 1 dangerous, : 1 date : 1 dating : 1 daughter : 2 days : 1 decided : 3 declassified : 1 deeply : 1 deliberately : 1 delivered : 1 descent, : 1 designer : 1 designer. : 1 desktop : 1 develop : 1 development : 3 diagnosed : 1 did : 5 died : 1 different : 1 different" : 1 difficult : 1 directors : 1 division : 1 docked : 1 doctoral : 1 don't : 1 dropped : 1 dropping : 1 during : 2 dying : 1 economics : 1 ectopic : 1 education, : 1 effects : 1 egg : 1 engaged : 1 engine-room : 1 enlightenment : 1 entrepreneur, : 1 event : 1 eventually : 3 everyone."[17] : 1 executive : 1 fame : 1 family : 4 family[18] : 1 farm : 2 father : 3 father, : 2 father.[12] : 1 feature : 1 felt : 3 few : 2 film, : 1 financial : 1 find : 1 first : 6 followed : 1 following : 1 for : 12 forbade : 1 forced : 1 former : 1 found : 1 foundation : 1 founder, : 1 frightened : 1 from : 2 full : 1 fully : 1 funded : 1 gained : 1 gave : 1 girl : 1 girlfriend, : 1 give : 1 giving : 1 go : 1 going : 1 grandparents : 1 graphical : 1 graphics : 1 graphics. : 1 grew : 4 had : 11 halted : 1 happy : 1 harsh, : 1 has : 2 have : 3 having : 1 he : 19

helped : 1 her : 5 herself : 1 high : 2 higher-education : 1 highly : 1 him : 6 him, : 1 him. : 1  
 him."[18] : 1 his : 15 hobby, : 1 household,[15] : 1 household.[11] : 1 housewife.[11] : 1 however, :  
 1 husband : 1 iMac, : 1 iPad. : 1 iPhone, : 1 iPod, : 1 iTunes : 2 if : 1 illegal : 1 immigrants, : 1  
 important : 1 in : 41 including : 1 indulged : 1 industrial : 1 industry : 2 initially : 1 initiate : 1  
 instead.[18] : 1 interface : 1 into : 1 introduced : 1 inventor, : 1 involve : 1 is : 1 it : 2 it's : 2 its : 2  
 jail : 1 joined : 1 just : 1 killed : 1 knew : 1 knowing, : 1 larger : 1 laser : 1 later : 3 later, : 1 latter : 1  
 law, : 1 leading : 1 leave : 1 leaving : 1 led : 1 left : 1 life : 1 life.[8] : 1 line : 1 lived : 1 long : 1  
 looking : 1 lot : 1 love : 2 loved : 1 machinist.[15] : 1 made : 3 magnate, : 1 majority : 1 man", : 1  
 marijuana : 1 markets. : 1 married : 2 marry : 1 marry.[12] : 1 mass-produced : 2 maternal : 1 matter  
 : 1 me : 2 me, : 2 me. : 1 member : 1 members : 1 merged : 1 merger, : 1 met : 1 microcomputer : 1  
 millionaire : 1 mind, : 1 mistake. : 1 modern : 1 months : 2 more."[12] : 1 most : 1 mother : 3  
 mother's : 1 mother, : 1 mothers, : 1 mouse-driven : 1 move : 1 moves, : 1 much : 2 my : 1 neither :  
 1 neuroendocrine : 1 new : 2 not : 8 noted : 1 notes : 1 nothing : 1 of : 41 off : 1 officer : 1 official :  
 2 on : 6 once : 1 one : 1 only : 2 onto : 1 option : 1 or : 1 original : 1 ostensible : 1 out : 3 out,[5] : 1  
 pancreatic : 1 papers.[12] : 1 parents : 4 parents" : 1 parents, : 1 parents."[19] : 1 part : 1 personal : 2  
 personality."[15] : 1 pioneers : 1 place."[17] : 1 placed : 2 platform : 1 platform, : 1 political : 2  
 possible : 1 potential : 1 power : 1 pregnancy, : 1 pregnant : 1 printer : 1 process : 1 produce : 1  
 products : 1 promised : 1 promptly : 1 publishing : 1 pursued : 1 put : 2 quietly : 1 raised : 2 raising :  
 1 ramifications : 1 really : 2 rebuilt : 1 recognized : 1 referred : 2 refused : 1 regard : 1 related : 1  
 relationship.[12] : 1 releasing : 1 replaced : 1 report : 1 reporter : 1 resemblance : 1 respiratory : 1  
 return : 1 revived : 1 revolution : 1 rise : 1 sadly, : 1 same : 1 saw : 1 say : 1 scared : 1 school, : 2  
 science.[11] : 1 seeking : 1 self-made : 1 sell : 1 series : 1 settled : 1 several : 1 shame : 1 shared : 1  
 shareholder : 1 she : 6 sheltered : 1 ship : 1 sign : 1 single : 1 sister, : 1 six : 1 so : 3 sometimes : 1  
 son : 2 specialized : 1 spent : 2 sperm : 2 spinout : 1 start : 1 state-of-the-art : 1 stated : 4 states : 1  
 stigma : 1 strong : 1 struggle, : 1 student : 1 study : 2 studying : 1 subject : 1 successful : 1 sudden :  
 1 suited : 1 summer : 1 support. : 1 take : 2 taking : 1 taking, : 1 tattoos, : 1 teaching : 1 telling : 1  
 ten : 1 that : 22 that[18] : 1 the : 66 their : 5 them : 3 then : 3 there : 2 they : 5 thing, : 1 things" : 1  
 this : 1 thought : 1 three : 1 through : 1 time : 2 time, : 1 time. : 1 to : 42 told : 3 too : 2 took : 2  
 tough : 1 tour : 1 traditional : 1 traveled : 2 tumor : 1 tumor. : 1 two : 3 tyrant, : 1 undergraduate : 1  
 unsuccessful : 1 unusual."[14] : 1 unwed : 1 up : 6 upped : 1 upset : 1 used : 1 user : 1 vector : 1  
 verge : 1 very : 1 visionaries : 1 visual : 1 want : 3 wanted : 3 war. : 1 was : 33 was, : 1 wasn't : 1  
 way : 1 we : 2 wealth : 1 wealthy."[18] : 1 wedlock : 1 well-educated, : 1 went : 1 were : 10 when :  
 4 where : 1 which : 2 while : 1 who : 5 whom : 1 widely : 1 wife : 2 with : 15 without : 1 women : 1  
 won : 1 work.[12][15] : 1 worked : 1 would : 6 year : 1 years : 2 young : 1

We realize another map reduce function in order to count the number\_of\_words in the text, and thus to display the right number.

## Question 2

the : 66  
 and : 53  
 a : 45  
 to : 42  
 in : 41

We use the question 1 and we just modify the order of the data by sorting them using their second arguments.

### Question 3

Apple : 11  
Clara : 8  
Jandali : 8  
Jobs's : 8  
Schieble : 8

We use again the data in "counts" but we filter and we keep only those who have more than five letter.

### Question 4

('60589', ('United States', 8145))  
('30594', ('France', 7799))  
('24449', ('Communes of France', 5740))  
('26539', ('Departments of France', 5299))  
('51359', ('Regions of France', 4064))  
('23683', ('City', 3832))  
('52174', ('Romania', 3527))  
('20409', ('Category:Rivers in Romania', 2978))  
('59931', ('Tributary', 2799))  
('28563', ('England', 2277))

We don't keep the first element of the line, because it doesn't count in the in-degree.

We count the in-degree of every pages.

The first element of the couple that we creat with map (lambda s : (s.split(" ")[0],(" ").join(s.split(" ")[1:])))) is the id of the page, and the second is the name of the page.

We join with the id, so we have (the id, (the name of the website, his in-dregree)).

Finally we sort the in-degrees.