Using Python for extracting most common words from a text

Simple Python script without the use of heavy text processing libraries to extract most common words from a corpus.

https://towardsdatascience.com/very-simple-python-script-for-extracting-most-common-words-from-a-story-1e3570d0b9d0 (https://towardsdatascience.com/very-simple-python-script-for-extracting-most-common-words-from-a-story-1e3570d0b9d0)

Interview in the newspaper:

https://www.publico.pt/2017/10/07/tecnologia/entrevista/o-grande-risco-da-inteligencia-artificial-sao-maquinas-demasiado-estupidas-1787954 (https://www.publico.pt/2017/10/07/tecnologia/entrevista/o-grande-risco-da-inteligencia-artificial-sao-maquinas-demasiado-estupidas-1787954)

Developed at 7-10-2018 by MRobalinho

In [1]:

```
# Some Documentation:
# https://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html
# https://medium.com/@MarutiTech/which-are-the-popular-languages-for-data-science-8e67fb5ef1ff
# https://www.kdnuggets.com/2017/01/most-popular-language-machine-learning-data-science.html
```

In [2]:

```
# http://www.storybench.org/getting-started-with-python-and-jupyter-notebooks-for-data-analysis/
```

In [3]:

```
# Libraries
import collections
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

In [4]:

```
# Path to acess the file
path = 'ml/count_words/'
```

In [5]:

```
# Read input file, note the encoding is specified here
# It may be different in your text file
my_text = 'Entrevista_Pedro_Domingos.txt'

file = open(path + my_text, encoding="utf8")
text_file = file.read()
```

In [6]:

```
# My text
#text_file
```

In [7]:

```
from pprint import pprint
```

In [8]:

```
# print my text file to analyze
pprint(text_file, width=100)
```

```
('“O grande risco da inteligência artificial são máquinas demasiado estúpi
das” \n'
 '\n'
 '“O grande risco da inteligência\n'
 'artificial são máquinas\n'
 'demasiado estúpidas”\n'
 'O académico Pedro Domingos, autor de A Revolução do\n'
 'Algoritmo Mestre, antevê um mundo em que os computadores\n'
 'poderão ser capazes de aprender tudo.\n'
 'Os computadores já são capazes de\n'
 'aprender: analisam enormes quantidades\n'
 'de dados e aprendem o que cada um de nós\n'
 'quer comprar, o tipo de pessoa por quem\n'
 'nos sentimos atraídos, quais os empregos\n'
 'que podem ser um bom passo na carreira,\n'
 'quais de nós podem vir a ser terroristas.\n'
 'O livro A Revolução do Algoritmo Mestre\n'
 '(editora Manuscrito), do académico e\n'
 'investigador Pedro Domingos, descreve\n'
 'como estes algoritmos são usados por\n'
 'empresas como a Amazon, o Google e o\n'
 'Facebook, e quais os impactos que já estão\n'
 'a ter no mundo. E avança para a busca de\n'
 'um algoritmo-mestre, um sistema de\n'
 'aprendizagem universal capaz de deduzir\n'
 'todo o conhecimento, desde que para isso\n'
 'seja “alimentado” com a informação\n'
 'necessária. Um algoritmo-mestre seria\n'
 'capaz tanto de manter o nosso email livre\n'
 'de spam como de descobrir curas\n'
 '(personalizadas) para cancros. Porém, um\n'
 'mundo de máquinas cada vez mais\n'
 'inteligentes não está isento de riscos. Para\n'
 'que ferramentas muito poderosas não\n'
 'estejam escondidas dentro de empresas, o\n'
 'autor defende que os utilizadores devem\n'
 'conhecer como estes algoritmos funcionam\n'
 'e ter uma palavra a dizer na forma como\n'
 'são integrados na sociedade.\n'
 'Pedro Domingos é professor de Ciências da\n'
 'Computação na Universidade de\n'
 'Washington. Conversou com o PÚBLICO a\n'
 'propósito do lançamento da versão\n'
 'portuguesa do livro, que decorreu nesta\n'
 'sexta-feira. O original foi publicado há dois\n'
 'anos, em inglês, e é – como se faz questão\n'
 'de dizer na capa – recomendado por Bill\n'
 'Gates.\n'
 'Por que é importante para o\n'
 'utilizador conhecer a forma como os\n'
 'algoritmos de aprendizagem\n'
 'funcionam? Já sabemos usar o\n'
 'Netflix, a Amazon, o Google.\n'
 'Sabemos usá-los para obter o objectivo\n'
 'imediato. Mas, ao mesmo tempo que\n'
 'estamos a atingir esse objectivo imediato,\n'
 '\n'
 'estamos a ensinar aos computadores aquilo\n'
 'que queremos. É preciso as pessoas\n'
 'estarem conscientes disso, para que essa\n'
 'aprendizagem dos computadores as sirva,\n'
```

```
'para ser eu a decidir e não outros a\n'
'decidirem por mim. A Internet criou este\n'
'mundo de escolha infinita, no qual em vez\n'
'de escolher de dez mil livros numa livraria,\n'
'posso escolher de dez milhões na Amazon.\n'
'Mas quem faz essa escolha? A Amazon. O\n'
'importante é que essa escolha seja feita da\n'
'maneira que eu faria se estivesse a ler os\n'
'livros um a um. A aprendizagem hoje ainda\n'
'é muito imperfeita. Os sistemas fazem\n'
'muitos disparates, recomendações que não\n'
'fazem sentido.\n'
'Estes sistemas são tão abertos que o\n'
'utilizador final pode fazer essa\n'
'diferença? A Amazon e o Netflix\n'
'terão interesse em canalizar o\n'
'utilizador para determinadas\n'
'escolhas.\n'
'Há dois aspectos. Um é o que as pessoas\n'
'podem fazer com os algoritmos como eles\n'
'existem hoje. Podem ensinar ao Google\n'
'aquilo de que gostam e ter resultados de\n'
'pesquisa que são de facto aqueles que\n'
'queriam. A longo prazo o aspecto mais\n'
'\n'
'importante é exigirmos que a caixa preta se\n'
'abra. Não precisamos de ter acesso ao\n'
'motor do carro. Mas precisamos de ter\n'
'acesso ao volante e aos pedais.\n'
'Acha que isso é possível?\n'
'A partir do momento em que há exigência\n'
'por parte das pessoas, o Google e a Amazon\n'
'fazem isso, ou aparecem outras empresas\n'
'que fazem. O papel mais importante é do\n'
'indivíduo, do consumidor e cidadão.\n'
'Hoje há outra grande balança de poder que está muito\n'
'desequilibrada: é o poder do conhecimento.\n'
'Sugere a criação de sindicatos de\n'
'dados para lidar com essas\n'
'empresas. É uma sugestão concreta\n'
'ou uma provocação?\n'
'Concreta. Precisamos de ter modelos\n'
'nossos que sejam o mais inteligentes e o\n'
'mais completos possível. O ideal é ter um\n'
'modelo meu que me conhece tão bem\n'
'como o meu melhor amigo. Este modelo\n'
'faz o meu papel no ciberespaço: lê esses\n'
'dez milhões de livros para escolher os dez\n'
'de que eu vou gostar. Quero que esse\n'
'modelo exista, mas não quero que esteja\n'
'sob o controlo do Google ou do Facebook.\n'
'Uma das opções que ponho é o banco de\n'
'dados: algo que é para os nossos dados o\n'
'que o banco é para o dinheiro. O banco não\n'
'foge com o nosso dinheiro. Outra\n'
'alternativa é algo como um sindicato de\n'
'dados. Os sindicatos surgiram para\n'
'equilibrar a balança de poder entre os\n'
'patrões e os empregados. Hoje há outra\n'
'grande balança de poder que está muito\n'
'desequilibrada: é o poder do\n'
```

```
'conhecimento. O Google e o Facebook cada\n'
'vez sabem mais e nós continuamos a não\n'
'saber muito sobre eles.\n'
'Como se chega a um sindicato\n'
'desses? Os sindicatos de\n'
'trabalhadores foram criados porque\n'
'o desequilíbrio se reflectia de forma\n'
'palpável na vida dos trabalhadores.\n'
'Os algoritmos tornam a vida mais\n'
'confortável. E as pessoas nem sequer\n'
'lêem as políticas de privacidade.\n'
'É preciso as coisas tornarem-se mais\n'
'palpáveis e os especialistas da área devem\n'
'ajudar as pessoas a fazerem essa tomada de\n'
'consciência. Por outro lado, mais cedo ou\n'
'mais tarde, vão acontecer coisas más e as\n'
'pessoas vão dar-se conta de que algo está\n'
'errado.\n'
'No livro apresenta a aprendizagem\n'
'automática como uma área especial,\n'
'apesar de haver outras áreas de\n'
'inteligência artificial. Porquê?\n'
'A era da informação tinha inicialmente\n'
'algoritmos programados por seres\n'
'humanos. Se quisesse que um algoritmo\n'
'fizesse diagnóstico médico tinha de\n'
'explicar ponto a ponto como o fazer. Isso já\n'
'nos deu grandes coisas. Os\n'
'programadores passaram a ser o factor\n'
'limitador. Com a aprendizagem, os\n'
'\n'
'sabemos como programar. Um carro sem\n'
'condutor: sabemos guiar, mas não\n'
'sabemos como o programar. Ele aprende\n'
'observando o vídeo da estrada e as acções\n'
'das pessoas no volante e nos pedais.\n'
'Este conceito do algoritmo-mestre é\n'
'o de que, com uma quantidade\n'
'suficiente de dados, o computador\n'
'aprende o que quer que seja. Mas\n'
'como é que vamos ensinar ética a um\n'
'carro ou a uma arma autónoma? Que\n'
'tipo de dados podemos dar para uma\n'
'máquina aprender a tomar decisões\n'
'morais?\n'
'A primeira coisa é dizermos ao carro quais\n'
'são as regras. Pode ser útil, mas não chega.\n'
'A outra hipótese é a da aprendizagem\n'
'automática: os carros, ou outros robôs,\n'
'aprenderem observando as pessoas.\n'
'Observam se as pessoas num caso\n'
'atropelam e se noutro caso caem de uma\n'
'ponte. Foi feito um inquérito a perguntar\n'
'se um carro autónomo deve atirar-se para\n'
'o rio para salvar pessoas. As pessoas\n'
'disseram que sim. Mas se a pessoa estiver\n'
'dentro carro, aí diz que não. Se os\n'
'algoritmos quiserem aprender connosco\n'
'vão ficar bastante confusos, porque nós\n'
'próprios não somos coerentes no nosso\n'
'comportamento ético. A aprendizagem\n'
```

```
'automática vai-nos obrigar a enfrentar\n'
'essas questões. No caso dos carros,\n'
'\n'
'pessoas. Mas é possível que as pessoas\n'
'depois vão comprar software pirata e\n'
'programem o carro de outra forma.\n'
'Muitas das vezes que fala do\n'
'algoritmo-mestre parece referir-se\n'
'mais à descoberta de algo que já\n'
'existe do que a uma invenção.\n'
'Na ciência em geral fala-se de descobertas:\n'
'descobrimos as leis da física. Na\n'
'tecnologia, fala-se de invenção: inventámos\n'
'o computador. Por um lado, podemos\n'
'inventar algoritmos, mas por outro lado\n'
'estamos a descobrir leis da aprendizagem,\n'
'a que os seres humanos também\n'
'obedecem. Os algoritmos de aprendizagem\n'
'que temos hoje já são algoritmos-mestres\n'
'no sentido em que o mesmo algoritmo\n'
'serve para coisas diferentes, ao contrário\n'
'dos algoritmos tradicionais: o algoritmo\n'
'que joga xadrez só joga xadrez. Mas os\n'
'diferentes paradigmas de aprendizagem,\n'
'baseados na evolução ou no cérebro, só são\n'
'capazes de aprender algumas coisas. O que\n'
'precisamos para um algoritmo-mestre é\n'
'que seja capaz de aprender todas essas\n'
'coisas diferentes.\n'
'Isso significa criar algoritmos que\n'
'aprendem e que têm processos que\n'
'nós não sabemos bem quais são.\n'
'Apenas vemos o resultado. Não é um\n'
'problema?\n'
'É a diferença entre a tecnologia actual e a\n'
'de há décadas. Antes compreendíamos\n'
'completamente a tecnologia. Hoje um\n'
'carro está cheio de computadores. E não há\n'
'\n'
'provavelmente o que vai haver são leis, que\n'
'dizem, por exemplo, que o carro deve\n'
'sacrificar o condutor se isso salvar mais\n'
'\n'
'computadores programam-se a si próprios.\n'
'Há coisas que queremos fazer que nem\n'
'\n'
'nenhuma pessoa no fabricante que saiba\n'
'como funcionam todas as partes de um\n'
'carro. Como o algoritmo é aprendido\n'
'[resulta de aprendizagem], estamos apenas\n'
'a julgar pelo exterior se ele está a fazer as\n'
'coisas certas ou não. Sempre vivemos num\n'
'mundo que só compreendemos\n'
'parcialmente. Mas a tecnologia está sob\n'
'nosso controlo. Pelo facto de não serem\n'
'completamente compreensíveis não\n'
'significa que não possam ser extensões\n'
'nossas.\n'
'Mas há riscos de danos colaterais,\n'
'mesmo que o algoritmo esteja a fazer\n'
'bem o que é suposto fazer.\n'
```

```
'Recentemente, o Facebook sugeria\n'
'publicidade dirigida para grupos\n'
'anti-semitas. Qual é a dimensão\n'
'deste risco?\n'
'É grande. Aparece em muitas áreas\n'
'diferentes. O problema que o Facebook\n'
'ilustra bem é que os algoritmos estão a\n'
'usar inteligência para atingir certos\n'
'objectivos. Mas esses algoritmos não\n'
'compreendem uma série de outras coisas\n'
'importantes. Não têm senso comum. Pode\n'
'dar resultados maus.\n'
'Este algoritmo-mestre não\n'
'conheceria também tudo do mundo.\n'
'Mesmo uma solução destas não está\n'
'isenta de efeitos nocivos colaterais.\n'
'Não, por várias razões. O algoritmo-mestre\n'
'é apenas um algoritmo de aprendizagem.\n'
'Depende dos dados. Se lhe dermos dados\n'
'que não prestam, não faz milagres. Mas\n'
'\n'
'além dos dados, recebe os objectivos. Se o\n'
'objectivo for mau, o algoritmo vai fazer\n'
'mal. O objectivo do Facebook é maximizar\n'
'o envolvimento das pessoas. Quando o\n'
'algoritmo está a maximizar esse objectivo,\n'
'não está a maximizar as notícias\n'
'verdadeiras, portanto escolhe as notícias\n'
'más, que muitas vezes as pessoas lêem e\n'
'comentam mais.\n'
'O algoritmo-mestre, se o atingirmos, é demasiado\n'
'poderoso e demasiado importante para estar nas mãos\n'
'de uma empresa.\n'
'Pedro Domingos, investigador\n'
'Os algoritmos hoje são coisas\n'
'protegidas. Chegando ao algoritmomestre,\n'
'temos pelo menos dois\n'
'cenários: ou alguém o disponibiliza\n'
'livremente, ou fica fechado numa\n'
'empresa. Este segundo cenário seria\n'
'preocupante?\n'
'Seria. O algoritmo-mestre, se o atingirmos,\n'
'é demasiado poderoso e demasiado\n'
'importante para estar nas mãos de uma\n'
'empresa. Na informática as coisas não são\n'
'muito patenteáveis. Há algoritmos que\n'
'foram patentados por empresas, mas é\n'
'possível fazer-se uma variação que já não\n'
'está coberta pela patente. E há uma\n'
'tradição muito importante de software\n'
'open source.\n'
'Como é que um especialista vê o\n'
'circo mediático em torno dos riscos\n'
'da inteligência artificial?\n'
'\n'
'Há riscos verdadeiros e riscos imaginários.\n'
'O circo maior é em torno dos riscos\n'
'imaginários e distrai as pessoas dos\n'
'verdadeiros. Um dos imaginários é esta\n'
'ideia de que as máquinas se revoltam e\n'
'tentam controlar o mundo. Outro risco\n'
```

```
'muito maior é o risco das máquinas\n'
'incompetentes: máquinas que tomam\n'
'decisões erradas por não perceberem\n'
'melhor. Ironicamente, o grande risco da\n'
'inteligência artificial não são máquinas\n'
'demasiado inteligentes, são máquinas\n'
'demasiado estúpidas. As máquinas já\n'
'tomam uma série de decisões muito\n'
'importantes: que candidatos a empregos\n'
'são entrevistados por empresas, quem são\n'
'os potenciais criminosos ou terroristas,\n'
'quem se recomenda a outra pessoa para\n'
'que saiam juntos.\n'
'Descreve no livro um cenário em que\n'
'há um modelo virtual das pessoas,\n'
'que vai à entrevista de emprego por\n'
'elas. Esse modelo não pode também\n'
'trabalhar pela pessoa?\n'
'Se o modelo pode fazer tudo, as pessoas\n'
'podem ir de férias. É uma possibilidade no\n'
'futuro e a questão será então como\n'
'distribuir a riqueza. Mas a curto e médio\n'
'prazo, os modelos são uma versão muito\n'
'imperfeita das pessoas.\n'
'A longo prazo acha que vamos todos\n'
'de férias?\n'
'A muito longo prazo – décadas ou talvez\n'
'centenas de anos – é possível que a\n'
'inteligência artificial e os robôs sejam\n'
'\n'
'capazes de fazer tudo melhor do que os\n'
'seres humanos. Aí, os seres humanos vão\n'
'usufruir do mundo tecnológico da mesma\n'
'forma que usufruíam dos frutos das\n'
'árvores. Mas como é que esses frutos são\n'
'divididos? São divididos por quem controla\n'
'as empresas ou distribuídos por todos nós?\n'
'Enquanto a maioria das pessoas quiser que\n'
'os frutos sejam distribuídos, eles serão\n'
'distribuídos. É por isso que, a longo prazo,\n'
'o nosso voto é mais importante do que o\n'
'nosso emprego.\n'
'Estas discussões sobre inteligência\n'
'artificial desembocam quase sempre\n'
'em questões muito fundamentais.\n'
'Desde a desigualdade até questões\n'
'sobre o que é ser humano. Isso faz\n'
'com que as pessoas da área se sintam\n'
'como os programadores do futuro?\n'
'Descreve no livro os programadores\n'
'como deuses menores.\n'
'A inteligência artificial é diferente de\n'
'muitas outras áreas da tecnologia\n'
'precisamente porque é muito fundamental.\n'
'Uma coisa é automatizar o trabalho\n'
'manual, outra coisa é automatizar a\n'
'inteligência, que é aquilo de mais profundo\n'
'e mais único que temos. A inteligência\n'
'artificial dá-nos todos estes poderes. Mas\n'
'não são os programadores nem os\n'
'especialistas que vão decidir como estes\n'
```

```
 'poderes vão ser utilizados. Decisões éticas\n'
 'têm de ser tomadas pelos indivíduos, pela\n'
 'sociedade.\n'
 '\n')
```

In [9]:

```
# Stopwords - File with words we don't need count
# read one word each line of stopwords
stopwords = set(line.strip() for line in open(path + 'stopwords_pt.txt'))
stopwords = stopwords.union(set(['mr','mrs','one','two','said']))
```

In [10]:

```
# Print the words i dont't want to count
print (stopwords)
```

```
{'das', 'algm', 'faz', 'isso', 'o', 'por', 'no', 'seja', '3', 'lhe', 'pedr
o', 'estes', 'vez', '8', 'ele', 'aqueles', '9', 'quais', 'na', 'seria', 'a
s', 'para', 'grande', 'alguns', '4', 'em', 'algo', 'fazem', 'said', 'que
m', 'uma', 'nosso', 'que', 'vão', 'há', 'mr', 'algumas', 'aí', 'cada', 'pe
la', '5', 'eles', 'one', 'um', 'da', 'muitas', 'alguem', 'noutro', 'mesm
o', 'ao', 'como', '1', 'tudo', 'numa', 'se', 'este', 'nas', 'esses', 'el
a', 'num', 'estas', 'mas', 'alguma', 'está', 'ter', 'mais', 'ser', 'form
a', 'mrs', 'aos', 'podem', '2', 'elas', 'nós', 'outras', 'outro', 'me', 'o
utra', 'two', 'mim', 'meu', 'muito', 'nos', 'qual', 'mesma', 'com', 'não',
'é', '6', '0', 'esta', 'ainda', 'serem', 'ou', 'apesar', '7', 'já', 'pod
e', 'de', 'coisas', 'essas', 'pelos', 'dez', 'os', 'essa', 'do', 'outros',
'a', 'quando', 'apenas', 'dos', 'nem', 'aquilo', 'esse', 'são', 'e', 'pel
o', 'bem'}
```

Count Text Words

In [11]:

```
# Instantiate a dictionary, and for every word in the file,
# Add to the dictionary if it doesn't exist. If it does, increase the count.
wordcount = {}
```

In [12]:

```
# To eliminate duplicates, remember to split by punctuation, and use case demiliters.
for word in text_file.lower().split():
    word = word.replace(".","")
    word = word.replace(",","")
    word = word.replace(":","")
    word = word.replace("\"","")
    word = word.replace("!","")
    word = word.replace("â€œ","")
    word = word.replace("â€˜","")
    word = word.replace("*","")
    if word not in stopwords:
        if word not in wordcount:
            wordcount[word] = 1
        else:
            wordcount[word] += 1
```

In [13]:

```
pprint(wordcount)
```

```
{'(editora': 1,
 '(personalizadas)': 1,
 '[resulta': 1,
 'abertos': 1,
 'abra': 1,
 'académico': 2,
 'acesso': 2,
 'acha': 2,
 'acontecer': 1,
 'actual': 1,
 'acções': 1,
 'ajudar': 1,
 'algoritmo': 10,
 'algoritmo-mestre': 9,
 'algoritmomestre': 1,
 'algoritmos': 15,
 'algoritmos-mestres': 1,
 'alguém': 1,
 'alternativa': 1,
 'além': 1,
 'amazon': 6,
 'amigo': 1,
 'analisam': 1,
 'anos': 2,
 'antes': 1,
 'antevê': 1,
 'anti-semitas': 1,
 'aparece': 1,
 'aparecem': 1,
 'aprende': 2,
 'aprendem': 2,
 'aprender': 6,
 'aprenderem': 1,
 'aprendido': 1,
 'aprendizagem': 12,
 'aprendizagem]': 1,
 'apresenta': 1,
 'arma': 1,
 'artificial': 8,
 'artificial?': 1,
 'aspecto': 1,
 'aspectos': 1,
 'atingir': 2,
 'atingirmos': 2,
 'atirar-se': 1,
 'atraídos': 1,
 'atropelam': 1,
 'até': 1,
 'automatizar': 2,
 'automática': 3,
 'autor': 2,
 'autónoma?': 1,
 'autónomo': 1,
 'avança': 1,
 'balança': 3,
 'banco': 3,
 'baseados': 1,
 'bastante': 1,
 'bill': 1,
 'bom': 1,
 'busca': 1,
```

```
    'caem': 1,
    'caixa': 1,
    'canalizar': 1,
    'cancros': 1,
    'candidatos': 1,
    'capa': 1,
    'capaz': 3,
    'capazes': 4,
    'carreira': 1,
    'carro': 10,
    'carros': 2,
    'caso': 3,
    'cedo': 1,
    'centenas': 1,
    'cenário': 2,
    'cenários': 1,
    'certas': 1,
    'certos': 1,
    'chega': 2,
    'chegando': 1,
    'cheio': 1,
    'ciberespaço': 1,
    'cidadão': 1,
    'circo': 2,
    'ciência': 1,
    'ciências': 1,
    'coberta': 1,
    'coerentes': 1,
    'coisa': 3,
    'colaterais': 2,
    'comentam': 1,
    'completamente': 2,
    'completos': 1,
    'comportamento': 1,
    'comprar': 2,
    'compreendem': 1,
    'compreendemos': 1,
    'compreendíamos': 1,
    'compreensíveis': 1,
    'computador': 2,
    'computadores': 6,
    'computação': 1,
    'comum': 1,
    'conceito': 1,
    'concreta': 2,
    'condutor': 2,
    'confortável': 1,
    'confusos': 1,
    'conhece': 1,
    'conhecer': 2,
    'conheceria': 1,
    'conhecimento': 3,
    'connosco': 1,
    'conscientes': 1,
    'consciência': 1,
    'consumidor': 1,
    'conta': 1,
    'continuamos': 1,
    'controla': 1,
    'controlar': 1,
    'controlo': 2,
```

```
    'contrário': 1,
    'conversou': 1,
    'criados': 1,
    'criar': 1,
    'criação': 1,
    'criminosos': 1,
    'criou': 1,
    'curas': 1,
    'curto': 1,
    'cérebro': 1,
    'dados': 10,
    'danos': 1,
    'dar': 2,
    'dar-se': 1,
    'decidir': 2,
    'decidirem': 1,
    'decisões': 4,
    'decorreu': 1,
    'deduzir': 1,
    'defende': 1,
    'demasiado': 8,
    'dentro': 2,
    'depende': 1,
    'depois': 1,
    'dermos': 1,
    'descoberta': 1,
    'descobertas': 1,
    'descobrimos': 1,
    'descobrir': 2,
    'descreve': 3,
    'desde': 2,
    'desembocam': 1,
    'desequilibrada': 2,
    'desequilíbrio': 1,
    'desigualdade': 1,
    'desses?': 1,
    'destas': 1,
    'deste': 1,
    'determinadas': 1,
    'deu': 1,
    'deuses': 1,
    'deve': 2,
    'devem': 2,
    'diagnóstico': 1,
    'diferente': 1,
    'diferentes': 4,
    'diferença': 1,
    'diferença?': 1,
    'dimensão': 1,
    'dinheiro': 2,
    'dirigida': 1,
    'discussões': 1,
    'disparates': 1,
    'disponibiliza': 1,
    'disseram': 1,
    'disso': 1,
    'distrai': 1,
    'distribuir': 1,
    'distribuídos': 3,
    'divididos': 1,
    'divididos?': 1,
```

```
    'diz': 1,
    'dizem': 1,
    'dizer': 2,
    'dizermos': 1,
    'dois': 3,
    'domingos': 4,
    'dá-nos': 1,
    'décadas': 2,
    'efeitos': 1,
    'email': 1,
    'empregados': 1,
    'emprego': 2,
    'empregos': 2,
    'empresa': 3,
    'empresas': 7,
    'enfrentar': 1,
    'enormes': 1,
    'enquanto': 1,
    'ensinar': 3,
    'entre': 2,
    'entrevista': 1,
    'entrevistados': 1,
    'então': 1,
    'envolvimento': 1,
    'equilibrar': 1,
    'era': 1,
    'erradas': 1,
    'errado': 1,
    'escolha': 2,
    'escolha?': 1,
    'escolhas': 1,
    'escolhe': 1,
    'escolher': 3,
    'escondidas': 1,
    'especial': 1,
    'especialista': 1,
    'especialistas': 2,
    'estamos': 4,
    'estar': 2,
    'estarem': 1,
    'esteja': 2,
    'estejam': 1,
    'estiver': 1,
    'estivesse': 1,
    'estrada': 1,
    'estão': 2,
    'estúpidas': 1,
    'estúpidas”': 2,
    'eu': 3,
    'evolução': 1,
    'exemplo': 1,
    'exigirmos': 1,
    'exigência': 1,
    'exista': 1,
    'existe': 1,
    'existem': 1,
    'explicar': 1,
    'extensões': 1,
    'exterior': 1,
    'fabricante': 1,
    'facebook': 6,
```

```
    'facto': 2,
    'factor': 1,
    'fala': 1,
    'fala-se': 2,
    'faria': 1,
    'fazer': 10,
    'fazer-se': 1,
    'fazerem': 1,
    'fechado': 1,
    'feita': 1,
    'feito': 1,
    'ferramentas': 1,
    'fica': 1,
    'ficar': 1,
    'final': 1,
    'fizesse': 1,
    'foge': 1,
    'foi': 2,
    'for': 1,
    'foram': 2,
    'frutos': 3,
    'funcionam': 2,
    'funcionam?': 1,
    'fundamentais': 1,
    'fundamental': 1,
    'futuro': 1,
    'futuro?': 1,
    'férias': 1,
    'férias?': 1,
    'física': 1,
    'gates': 1,
    'geral': 1,
    'google': 6,
    'gostam': 1,
    'gostar': 1,
    'grandes': 1,
    'grupos': 1,
    'guiar': 1,
    'haver': 2,
    'hipótese': 1,
    'hoje': 7,
    'humano': 1,
    'humanos': 4,
    'ideal': 1,
    'ideia': 1,
    'ilustra': 1,
    'imaginários': 3,
    'imediato': 2,
    'impactos': 1,
    'imperfeita': 2,
    'importante': 8,
    'importantes': 2,
    'incompetentes': 1,
    'indivíduo': 1,
    'indivíduos': 1,
    'infinita': 1,
    'informação': 2,
    'informática': 1,
    'inglês': 1,
    'inicialmente': 1,
    'inquérito': 1,
```

```
    'integrados': 1,
    'inteligentes': 3,
    'inteligência': 11,
    'interesse': 1,
    'internet': 1,
    'inventar': 1,
    'inventámos': 1,
    'invenção': 2,
    'investigador': 2,
    'ir': 1,
    'ironicamente': 1,
    'isenta': 1,
    'isento': 1,
    'joga': 2,
    'julgar': 1,
    'juntos': 1,
    'lado': 3,
    'lançamento': 1,
    'leis': 3,
    'ler': 1,
    'lidar': 1,
    'limitador': 1,
    'livraria': 1,
    'livre': 1,
    'livremente': 1,
    'livro': 5,
    'livros': 3,
    'longo': 4,
    'lê': 1,
    'lêem': 2,
    'maior': 2,
    'maioria': 1,
    'mal': 1,
    'maneira': 1,
    'manter': 1,
    'manual': 1,
    'manuscrito)': 1,
    'mau': 1,
    'maus': 1,
    'maximizar': 3,
    'mediático': 1,
    'melhor': 3,
    'menores': 1,
    'menos': 1,
    'mestre': 2,
    'mil': 1,
    'milagres': 1,
    'milhões': 2,
    'modelo': 6,
    'modelos': 2,
    'momento': 1,
    'morais?': 1,
    'motor': 1,
    'muitos': 1,
    'mundo': 8,
    'máquina': 1,
    'máquinas': 9,
    'más': 2,
    'mãos': 2,
    'médico': 1,
    'médio': 1,
```

```
    'necessária': 1,
    'nenhuma': 1,
    'nesta': 1,
    'netflix': 2,
    'nocivos': 1,
    'nossas': 1,
    'nossos': 2,
    'notícias': 2,
    'nós?': 1,
    'obedecem': 1,
    'objectivo': 5,
    'objectivos': 2,
    'obrigar': 1,
    'observam': 1,
    'observando': 2,
    'obter': 1,
    'open': 1,
    'opções': 1,
    'original': 1,
    'palavra': 1,
    'palpáveis': 1,
    'palpável': 1,
    'papel': 2,
    'paradigmas': 1,
    'parcialmente': 1,
    'parece': 1,
    'parte': 1,
    'partes': 1,
    'partir': 1,
    'passaram': 1,
    'passo': 1,
    'patentados': 1,
    'patente': 1,
    'patenteáveis': 1,
    'patrões': 1,
    'pedais': 2,
    'perceberem': 1,
    'perguntar': 1,
    'pesquisa': 1,
    'pessoa': 4,
    'pessoa?': 1,
    'pessoas': 21,
    'pirata': 1,
    'podemos': 2,
    'poder': 5,
    'poderes': 2,
    'poderosas': 1,
    'poderoso': 2,
    'poderão': 1,
    'políticas': 1,
    'ponho': 1,
    'ponte': 1,
    'ponto': 2,
    'porque': 3,
    'porquê?': 1,
    'portanto': 1,
    'portuguesa': 1,
    'porém': 1,
    'possam': 1,
    'possibilidade': 1,
    'posso': 1,
```

```
    'possível': 4,
    'possível?': 1,
    'potenciais': 1,
    'prazo': 5,
    'precisamente': 1,
    'precisamos': 4,
    'preciso': 2,
    'preocupante?': 1,
    'prestam': 1,
    'preta': 1,
    'primeira': 1,
    'privacidade': 1,
    'problema': 1,
    'problema?': 1,
    'processos': 1,
    'professor': 1,
    'profundo': 1,
    'programadores': 4,
    'programados': 1,
    'programam-se': 1,
    'programar': 2,
    'programem': 1,
    'propósito': 1,
    'protegidas': 1,
    'provavelmente': 1,
    'provocação?': 1,
    'próprios': 2,
    'publicado': 1,
    'publicidade': 1,
    'público': 1,
    'quantidade': 1,
    'quantidades': 1,
    'quase': 1,
    'quer': 2,
    'queremos': 2,
    'queriam': 1,
    'quero': 2,
    'questão': 2,
    'questões': 3,
    'quiser': 1,
    'quiserem': 1,
    'quisesse': 1,
    'razões': 1,
    'recebe': 1,
    'recentemente': 1,
    'recomenda': 1,
    'recomendado': 1,
    'recomendações': 1,
    'referir-se': 1,
    'reflectia': 1,
    'regras': 1,
    'resultado': 1,
    'resultados': 2,
    'revoltam': 1,
    'revolução': 2,
    'rio': 1,
    'riqueza': 1,
    'risco': 5,
    'risco?': 1,
    'riscos': 6,
    'robôs': 2,
```

```
    'sabem': 1,
    'sabemos': 6,
    'saber': 1,
    'sacrificar': 1,
    'saiam': 1,
    'saiba': 1,
    'salvar': 2,
    'segundo': 1,
    'sejam': 3,
    'sem': 1,
    'sempre': 2,
    'senso': 1,
    'sentido': 2,
    'sentimos': 1,
    'sequer': 1,
    'seres': 4,
    'serve': 1,
    'será': 1,
    'serão': 1,
    'sexta-feira': 1,
    'si': 1,
    'significa': 2,
    'sim': 1,
    'sindicato': 2,
    'sindicatos': 3,
    'sintam': 1,
    'sirva': 1,
    'sistema': 1,
    'sistemas': 2,
    'sob': 2,
    'sobre': 3,
    'sociedade': 2,
    'software': 2,
    'solução': 1,
    'somos': 1,
    'source': 1,
    'spam': 1,
    'suficiente': 1,
    'sugere': 1,
    'sugeria': 1,
    'sugestão': 1,
    'suposto': 1,
    'surgiram': 1,
    'série': 2,
    'só': 3,
    'talvez': 1,
    'também': 3,
    'tanto': 1,
    'tarde': 1,
    'tecnologia': 5,
    'tecnológico': 1,
    'temos': 3,
    'tempo': 1,
    'tentam': 1,
    'terroristas': 2,
    'terão': 1,
    'tinha': 2,
    'tipo': 2,
    'todas': 2,
    'todo': 1,
    'todos': 3,
```

```
 'tomada': 1,
 'tomadas': 1,
 'tomam': 2,
 'tomar': 1,
 'tornam': 1,
 'tornarem-se': 1,
 'torno': 2,
 'trabalhadores': 2,
 'trabalhar': 1,
 'trabalho': 1,
 'tradicionais': 1,
 'tradição': 1,
 'tão': 2,
 'têm': 3,
 'universal': 1,
 'universidade': 1,
 'usados': 1,
 'usar': 2,
 'usufruir': 1,
 'usufruíam': 1,
 'usá-los': 1,
 'utilizador': 3,
 'utilizadores': 1,
 'utilizados': 1,
 'vai': 3,
 'vai-nos': 1,
 'vamos': 2,
 'variação': 1,
 'vemos': 1,
 'verdadeiras': 1,
 'verdadeiros': 2,
 'versão': 2,
 'vezes': 2,
 'vida': 2,
 'vir': 1,
 'virtual': 1,
 'vivemos': 1,
 'volante': 2,
 'voto': 1,
 'vou': 1,
 'várias': 1,
 'vê': 1,
 'vídeo': 1,
 'washington': 1,
 'xadrez': 2,
 'à': 2,
 'área': 3,
 'áreas': 3,
 'árvores': 1,
 'ética': 1,
 'éticas': 1,
 'ético': 1,
 'único': 1,
 'útil': 1,
 '–': 4,
 '“alimentado”': 1,
 '“o': 2}
```

In [14]:

```
# Print most common word
n_print = int(input("How many most common words to print: "))
print("\nOK. The {} most common words are as follows\n".format(n_print))
word_counter = collections.Counter(wordcount)
for word, count in word_counter.most_common(n_print):
    print(word, ": ", count)
```

```
How many most common words to print: 50

OK. The 50 most common words are as follows

pessoas :  21
algoritmos :  15
aprendizagem :  12
inteligência :  11
algoritmo :  10
dados :  10
fazer :  10
carro :  10
máquinas :  9
algoritmo-mestre :  9
artificial :  8
demasiado :  8
mundo :  8
importante :  8
empresas :  7
hoje :  7
computadores :  6
aprender :  6
amazon :  6
google :  6
facebook :  6
riscos :  6
sabemos :  6
modelo :  6
risco :  5
livro :  5
objectivo :  5
prazo :  5
poder :  5
tecnologia :  5
domingos :  4
capazes :  4
pessoa :  4
– :  4
estamos :  4
longo :  4
precisamos :  4
possível :  4
seres :  4
humanos :  4
programadores :  4
decisões :  4
diferentes :  4
descreve :  3
capaz :  3
conhecimento :  3
inteligentes :  3
dois :  3
utilizador :  3
ensinar :  3

In [15]:
```

```
# Close the file
file.close()
```

Using Bar Graph

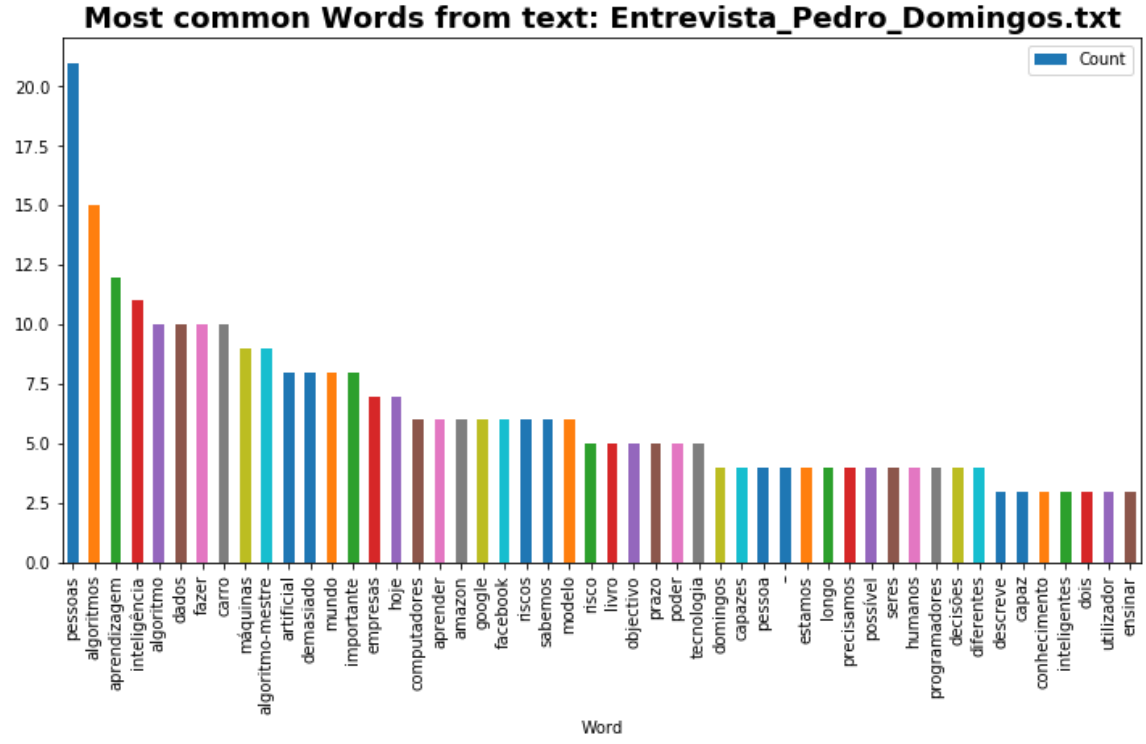
In [16]:

```
my_title = "Most common Words from text: "+my_text
```

In [17]:

```
# Create a data frame of the most common words
# Draw a bar chart
lst = word_counter.most_common(n_print)
df = pd.DataFrame(lst, columns = ['Word', 'Count'])

fig, ax1 = plt.subplots(figsize=(12,6))
ax1 = df.plot.bar(ax=ax1, x='Word',y='Count')
plt.title(my_title , fontdict={'size':18, 'weight': 'bold'});
```



In [18]:

```
df.head(10)
```

Out[18]:

	Word	Count
0	pessoas	21
1	algoritmos	15
2	aprendizagem	12
3	inteligência	11
4	algoritmo	10
5	dados	10
6	fazer	10
7	carro	10
8	máquinas	9
9	algoritmo-mestre	9

Using Word Cloud Graph

In [19]:

```
# To install the wordcloud use the bellow commands
# conda install -c conda-forge wordcloud
# https://www.datacamp.com/community/tutorials/wordcloud-python
# https://www.commonlounge.com/discussion/317a12109a634fc1aa44150ea806bbf3
# https://matplotlib.org/examples/color/colormaps_reference.html - Colormap to matplotl
ib

from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from matplotlib import cm
import numpy as np
```

In [20]:

```
wrds = df.Word

# WORDS without spaces
wrds =  df["Word"].str.replace(" ","")
wrds.head()
```

Out[20]:

```
0         pessoas
1      algoritmos
2    aprendizagem
3    inteligência
4       algoritmo
Name: Word, dtype: object
```

In [34]:

```
# Plot The WORDS in a Circle

x, y = np.ogrid[:300, :300]

mask = (x - 150) ** 2 + (y - 150) ** 2 > 130 ** 2
mask = 255 * mask.astype(int)

wc = WordCloud(background_color="white",  mask=mask, colormap=cm.plasma).generate(" ".j
oin(wrds))

plt.figure(figsize=(16,8))
plt.imshow(wc, interpolation="bilinear", origin='upper')
plt.axis("off")
#plt.tight_layout(pad=0)
plt.title(my_title , fontdict={'size':16, 'weight': 'bold'});
```

Most common Words from text: Entrevista_Pedro_Domingos.txt

In [22]:

```
# Plot The WORDS in a Frame

wc = WordCloud( background_color='white', colormap=cm.viridis, scale=5).generate(" ".jo
in(wrds))

plt.figure(figsize=(16,8))
plt.imshow(wc, interpolation="bilinear", origin='upper')
plt.axis("off")
plt.title(my_title , fontdict={'size':18, 'weight': 'bold'});
```

Using Pie Graph

In [23]:

```
from pandas.tools.plotting import table
```

In [26]:

```
# Obtain first 20 words
df1 = df.head(20)
df1.head()
```

Out[26]:

	Word	Count
0	pessoas	21
1	algoritmos	15
2	aprendizagem	12
3	inteligência	11
4	algoritmo	10

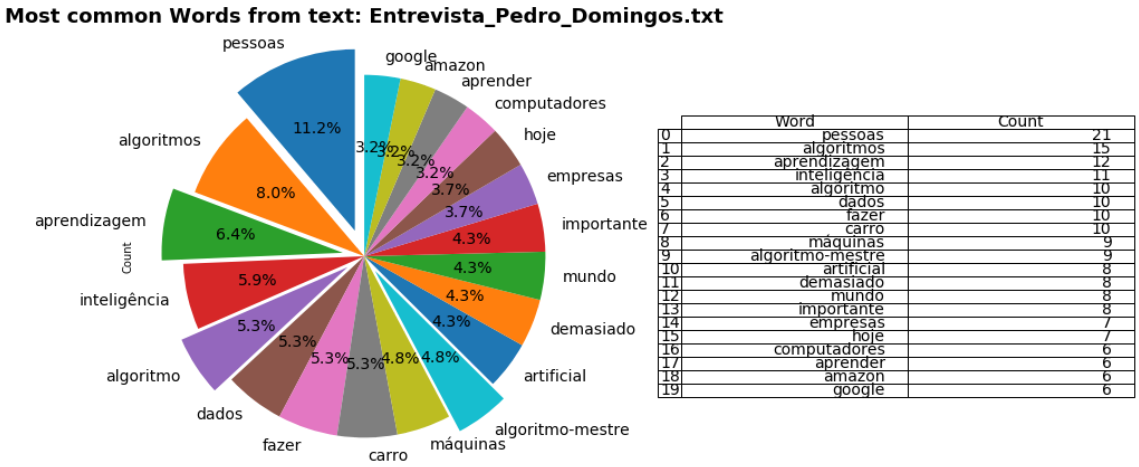
In [31]:

```
# First Graph - Plot Pie Graph
plt.figure(figsize=(16,10))
# plot chart
ax1 = plt.subplot(121, aspect='equal')
df1.plot(kind='pie', y = 'Count', ax=ax1, autopct='%1.1f%%',
         startangle=90, shadow=False,
         labels=df['Word'],   # Labels
         explode=(0.15, 0, 0.12, 0, 0.11,0,0,0,0,0.10,0,0,0,0,0,0,0,0,0,0), # draw a po
rtion
         legend = False, fontsize=14)

plt.title(my_title , fontdict={'size':18, 'weight': 'bold'});

# Second Graph - plot table
ax2 = plt.subplot(122)
plt.axis('off')
tbl = table(ax2, df1, loc='center')
tbl.auto_set_font_size(False)
tbl.set_fontsize(14)
plt.show()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:15: Futur
eWarning: 'pandas.tools.plotting.table' is deprecated, import 'pandas.plot
ting.table' instead.
  from ipykernel import kernelapp as app
```



	Word	Count
0	pessoas	21
1	algoritmos	15
2	aprendizagem	12
3	inteligência	11
4	algoritmo	10
5	dados	10
6	fazer	10
7	carro	10
8	máquinas	9
9	algoritmo-mestre	9
10	artificial	8
11	demasiado	8
12	mundo	8
13	importante	8
14	empresas	7
15	hoje	7
16	computadores	6
17	aprender	6
18	amazon	6
19	google	6

In [ ]:

```

```

In [ ]:

```

```