

1. 更新了字符串相似度比较算法，缩写与公司全名匹配的时候有相似度输出，使数据匹配更具有依据：

Jaro Distance是两个单词之间由一个转换为另一个所需的单字符转换的最小数量。

J-W距离 (Jaro-Winker distance)是由William E. Winkler在1990年提出的Jaro Distance度量标准的一种变体。

两者的区别在于：Jaro-Winkler Distance通过前缀因子 P 使Jaro Distance相同时共同前缀长度 L 越大的相似度越高。

如：

```
In [26]: jellyfish.jaro_winkler_similarity('AIR T INC',
      'OKASAN SECURITIES GROUP INC')
Out[26]: 0.6296296296296297

In [27]: jellyfish.jaro_winkler_similarity('CITIGROUP
      INC', 'OKASAN SECURITIES GROUP INC')
Out[27]: 0.527065527065527

In [28]: jellyfish.jaro_winkler_similarity('U.S.
      Government', 'ASA GOLD AND PRECIOUS METALS')
Out[28]: 0.4357142857142857

In [29]: jellyfish.jaro_winkler_similarity('China
      National Gold', 'ASA GOLD AND PRECIOUS METALS')
Out[29]: 0.36779448621553884
```

2. 改良后优点：就目前看来，经过目测匹配的数据可用度比之前的高了，准确率也更高。

	A	B	C	D	E	F	G	H
1	Upstream key	customer abbreviation	Full name	similarity	Full name	similarity	Full name	similarity
2	100644	Civil Infrastructure	COIL	0.535087719	CETIS	0.501754386	Anli International	0.496832359
3	026753	10 Customers	P10 INC	0.567099567	Notore Chemical In	0.515151515	U10 SA	0.505050505
4	154758	Managed Medicare	MMTC	0.544444444	Notore Chemical In	0.537037037	Anli International	0.522222222
5	010863	Commercial - Natural Gas	Mayberry Jamaica	0.596078431	eCloudvalley Digit	0.501851852	Anli International	0.489814815
6	031453	Commercial	Notore Chemical In	0.688888889	CGG	0.477777778	Anli International	0.459259259
7	011340	Not Reported	Notore Chemical In	0.627465127	NSTAR	0.527272727	Anli International	0.510774411
8	004809	Store Branded Retail	Anli International	0.555555556	SOLIBRA	0.531746032	Mayberry Jamaica	0.50497977
9	019516	MIHomes Inc	M2I	0.622222222	GAMEWITH INC	0.577777778	MIHOMES INC	0.569230769
10	021739	Non-Pharmaceutical	Notore Chemical In	0.61459695	SNRP	0.539215686	eCloudvalley Digit	0.50497977
11	031775	International	Anli International	0.779202279	Yankey Engineering	0.487179487	BIF	0.47008647
12	106969	LP Building Products (LP)	LPP SA	0.55	FIPP	0.533333333	OTP BANK PLC	0.511111111
13	187766	7 Customers	Notore Chemical In	0.533333333	CGG	0.477777778	eCloudvalley Digit	0.457407407
14	001224	Residential	Anli International	0.57984608	RH	0.53030303	Yankey Engineering	0.515151515
15	135410	Not Reported	Notore Chemical In	0.627465127	NSTAR	0.527272727	Anli International	0.510774411
16	032345	Micromax	Anli International	0.513888889	Notore Chemical In	0.490740741	MARR	0.458333333
17	170375	Primerica Life Insurance Company	Yankey Engineering	0.533461047	LISI	0.522988506	EPOLIN INC	0.512643678
18	179583	Lasers, Sensors, and Other	LISI	0.531746032	Yeah1 Group Ord Sh	0.525132275	ELSI	0.498412698
19	156656	Commercial - Electric	eCloudvalley Digit	0.57037037	Mayberry Jamaica	0.506535948	Yankey Engineering	0.497354497
20	014793	Central and South America	Notore Chemical In	0.617197617	Anli International	0.552188552	CEM SA	0.545454545
21	007849	Interrupible	Anli International	0.561609687	Yeah1 Group Ord Sh	0.487464387	Notore Chemical In	0.487179487
22	033528	Not Reported	Notore Chemical In	0.627465127	NSTAR	0.527272727	Anli International	0.510774411
23	164506	Staples/Essendant	SPIE	0.541666667	Yankey Engineering	0.513888889	MAN SE	0.486111111
24	148950	Aerospace	eCloudvalley Digit	0.555555556	Notore Chemical In	0.555555556	CAS	0.481481481
25	037200	Primoris Renewable Energy Inc	Yankey Engineering	0.598290598	PIERREL	0.514652015	RH	0.512820513
26	064028	South Korea	Notore Chemical In	0.601058201	Yeah1 Group Ord Sh	0.533333333	SKF AB	0.511111111
27	037909	108 Customers	P10 INC	0.55952381	1000MERCIS	0.516666667	U10 SA	0.5
28	174700	CMA CGM S.A.	COMCM S.A.	0.905	COMARCH SA	0.76	CONDMAG SA	0.76
29	018315	Gentien Therapeutics, Inc	Notore Chemical In	0.577248677	Yankey Engineering	0.531746032	GL LTD	0.476190476
30	106687	3D Buoy	DB INC	0.555555556	DNB ASA	0.53968254	ADOBE INC	0.518518519
31	010000	Not Reported	Notore Chemical In	0.627465127	NSTAR	0.527272727	Anli International	0.510774411
32	065590	General Motors Corp	GMS INC	0.535014006	GUMI INC	0.517156863	YRGLM INC	0.503267974
33	005849	Costco Wholesale Corp	Notore Chemical In	0.560623782	COWEN INC	0.497076023	CON-WAY INC	0.476874003
34	179817	Green Bay Packaging	Mayberry Jamaica	0.575163399	Notore Chemical In	0.552598817	Anli International	0.50497977
35	291881	Cipher Pharmaceuticals Inc	Notore Chemical In	0.611111111	Anli International	0.513888889	CAP-XX	0.472222222
36	003275	Education	ONE	0.481481481	VEF	0.481481481	eCloudvalley Digit	0.477777778
37	328795	General Electric Co	Anli International	0.513344227	eCloudvalley Digit	0.50497977	AMGEN INC	0.503267974
38	271342	Total SA	TAM SA	0.642857143	TXM SA	0.642857143	TUPY SA	0.619047619

3. **代码缺点**：代码循环嵌套，且数据量太大，需要跑很久。
-

接下来的工作：

1. 将匹配好的缩写和方翔匹配好的缩写比对，确认两人是否匹配数据可用。
2. 如果遇到少量类似于 'Other' 或 'United states' 类似的不明确的缩写名，可能会将这些数据排除在最后的数据之外
3. 在确定缩写匹配的数据无误之后，才能进行上下游的比对
如果顺利的话，按我的预想是这周五之前完成的。