ΜΥΕ003 - ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2018

Εργασία: "Μηχανή αναζήτησης κριτικών χρηστών" Σχεδιασμός και υλοποίηση συστήματος αναζήτησης πληροφορίας εστιατορίων και κριτικών εστιατορίων.

Γκαβαρδίνας Όθωνας, ΑΜ: 2620 Μπουρλή Στυλιανή, ΑΜ: 2774

> Συλλογή δεδομένων

Αρχικά, κατεβάσαμε τη JSON μορφή του Yelp dataset. Από τη συλλογή αυτή χρησιμοποιήσαμε το αρχείο business.json και το αρχείο review.json. Από το πρώτο συλλέξαμε τα εστιατόρια που περιείχαν περισσότερες από 50 κριτικές και από το δεύτερο συλλέξαμε τις κριτικές γι' αυτά τα εστιατόρια. Με βάση τις παραπάνω πληροφορίες δημιουργήσαμε ένα αρχείο κειμένου για κάθε εστιατόριο, με περιεχόμενο το όνομα του εστιατορίου, τη διεύθυνσή του, την πόλη και το κράτος στο οποίο βρίσκεται, τα αστέρια που κατέχει, καθώς επίσης και τον αριθμό των κριτικών που περιέχει. Επιπλέον για κάθε εστιατόριο στο αντίστοιχο αρχείο που αφορά το εστιατόριο υπάρχουν οι κριτικές που το αφορούν και για κάθε μία το κείμενό της, ο χρήστης που έγραψε την κριτική, η ημερομηνία που δημοσιεύθηκε, τα αστέρια της, πόσο χρήσιμη , πόσο αστεία και πόσο cool είναι.

Στοιχεία από τη συλλογή δεδομένων:

Συνολικός αριθμός επιλεγμένων εστιατορίων: 14.000
 Συνολικός αριθμός επιλεγμένων κριτικών: 2.536.053
 Μέσος αριθμός κριτικών για τα εστιατόρια: 181
 Μικρότερος αριθμός κριτικών για τα εστιατόρια: 51
 Μεναλύτερος αριθμός κριτικών για τα εστιατόρια: 7361

> Δημιουργία ευρετηρίου

Για τη δημιουργία του ευρετηρίου χρησιμοποιήσαμε τη lucene-7.2.1. Πιο συγκεκριμένα, ορίσαμε αρχικά ένα Directory με όλα τα έγγραφα που φτιάξαμε από το Yelp dataset, έναν StandardAnalyzer, για να σπάμε το περιεχόμενο των εγγράφων σε tokens, και έναν IndexWriter, για να εισάγουμε δεδομένα στο ευρετήριο. Στη συνέχεια, διαβάσαμε ένα ένα τα δεδομένα από τα έγγραφα στο Directory και για κάθε έγγραφο φτιάξαμε ένα Document με τα Fields: όνομα αρχείου, όνομα εστιατορίου, διεύθυνση εστιατορίου, πόλη και κράτος, στο οποίο βρίσκεται το εστιατόριο, πλήθος αστεριών και πλήθος κριτικών που κατέχει το εστιατόριο, καθώς

και το κείμενο κάθε κριτικής, τον χρήστη που την έγραψε, τον αριθμό των αστεριών της , την ημερομηνία της πιο πρόσφατης κριτικής γι' αυτό το έγγραφο, καθώς και τον αριθμό της πιο χρήσιμης. Η δημιουργία των Fields έγινε με τέτοιο τρόπο, ώστε τα πεδία: όνομα εστιατορίου, διεύθυνση εστιατορίου, πόλη και κράτος, στο οποίο βρίσκεται το εστιατόριο, κείμενο κριτικών εστιατορίου, χρήστης, που έγραψε την κριτική και αστέρια κριτικής, να έχουν δυνατότητα αναζήτησης, ενώ τα πεδία: αριθμός κριτικών εστιατορίου, αριθμός αστεριών εστιατορίου, ημερομηνία δημοσίευσης κριτικής και χρησιμότητα κριτικής, να έχουν δυνατότητα αναδιάταξης.

Αναζήτηση – Αναδιάταξη – Αντιπροσωπευτικά αποτελέσματα

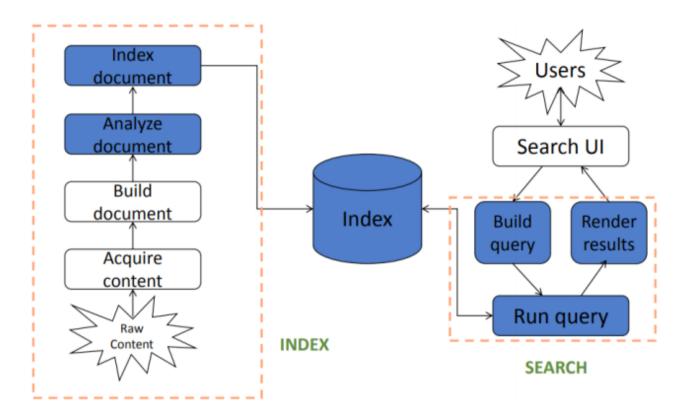
Για τη λειτουργία της αναζήτησης επιτρέψαμε στο χρήστη της μηχανής αναζήτησης να μπορεί να επιλέξει αν θέλει να κάνει αναζήτηση εστιατορίου ή κριτικής. Ανάλογα με την επιλογή αυτή να μπορεί ,έπειτα, να επιλέξει μεταξύ κάποιων πεδίων αναζήτησης, να εισάγει τους όρους αναζήτησης και να επιλέξει τρόπο διάταξης των αποτελεσμάτων.

Για τα εστιατόρια δίνεται επιλογή αναζήτησης με βάση τα πεδία: όνομα εστιατορίου, διεύθυνση εστιατορίου, πόλη και κράτος, στο οποίο βρίσκεται, κείμενο κριτικής, καθώς και συνδυασμό τοποθεσίας και κριτικής. Επίσης, δίνεται επιλογή διάταξης με βάση τον αριθμό των κριτικών και τον αριθμό των αστεριών, που κατέχει το εστιατόριο.

Για τις κριτικές δίνεται επιλογή αναζήτησης με βάση τα πεδία: κείμενο κριτικής, όνομα εστιατορίου, όνομα χρήστη που έγραψε την κριτική, και συνδυασμό ονόματος εστιατορίου και κειμένου κριτικής. Επίσης, δίνεται επιλογή διάταξης με βάση την πιο πρόσφατη κριτική και με βάση την πιο χρήσιμη κριτική. Συλλέγοντας αυτά τα δεδομένα, δημιουργήσαμε τη διαδικασία αναζήτησης και αναδιάταξης αντιπροσωπευτικών αποτελεσμάτων. Πιο συγκεκριμένα, ορίσαμε

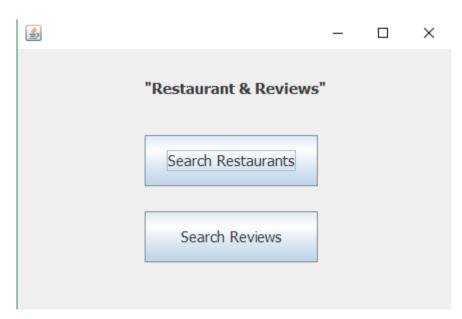
αρχικά έναν IndexReader και έναν IndexSearcher για αναζήτηση στο ευρετήριο, καθώς και έναν StandardAnalyzer, ίδιο με αυτό που χρησιμοποιήσαμε στη δημιουργία ευρετηρίου για να σπάσουμε το ερώτημα αναζήτησης του χρήστη σε tokens. Έπειτα, δημιουργήσαμε έναν MultiFieldQueryParser για τις περιπτώσεις που η αναζήτηση γίνεται σε πολλαπλά Fields και έναν QueryParser για την περίπτωση που η αναζήτηση γίνεται σε ένα μόνο Field. Ορίσαμε, επίσης, ένα Query για το ερώτημα του χρήστη. Σύμφωνα με το Field αναζήτησης και τον τρόπο διάταξης , που επέλεξε ο χρήστης, εκτελέσαμε την κατάλληλη διαδικασία αναζήτησης, συλλέγοντας τα 1000 πιο συναφή με το ερώτημα αποτελέσματα, κατάλληλα διατεταγμένα. Στη συνέχεια, αποθηκεύσαμε τα αποτελέσματα αυτά σε μία ScoreDoc λίστα και κάνοντας την κατάλληλη επεξεργασία κρατήσαμε τελικά 100 από αυτά, για τα οποία ισχύει ότι αν είναι εστιατόρια βρίσκονται σε διαφορετική τοποθεσία και αν είναι κριτικές, έχουν γραφτεί από διαφορετικούς χρήστες. Αυτό το κάναμε για να είναι τα αποτελέσματα πιο αντιπροσωπευτικά. Τέλος, παρουσιάσαμε τα αποτελέσματα αυτά στο χρήστη.

Όλη η παραπάνω διαδικασία παρουσιάζεται συνοπτικά στο παρακάτω σχεδιάγραμμα:



> Διεπαφή

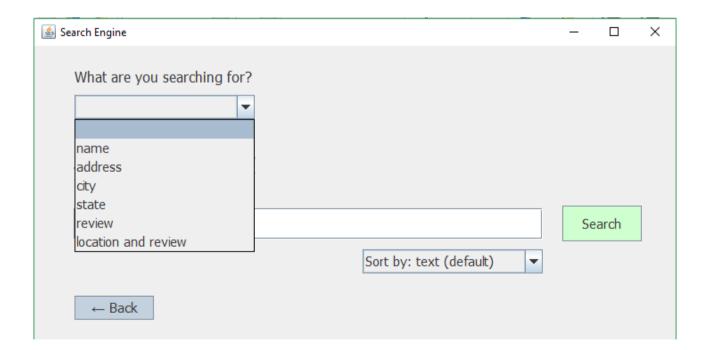
Για τη διεπαφή χρησιμοποιήσαμε το Swing Gui της Java. Αρχικά, εμφανίζεται στο χρήστη ένα παράθυρο με τις επιλογές: Search Restaurants, για αναζήτηση εστιατορίων και Search Reviews, για αναζήτηση κριτικών εστιατορίων.



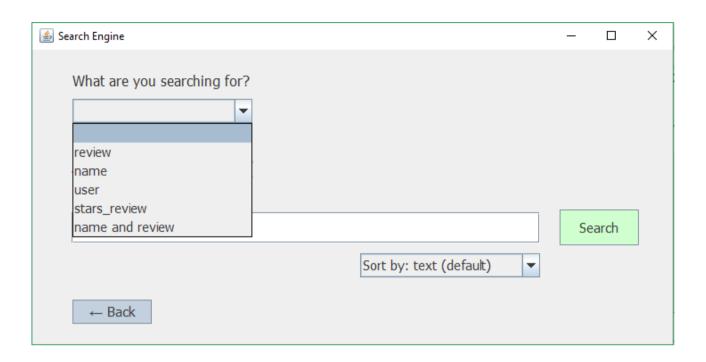
Αφού ο χρήστης επιλέξει την αναζήτηση που θέλει να πραγματοποιήσει, εμφανίζεται το κύριο παράθυρο της μηχανής αναζήτησης.



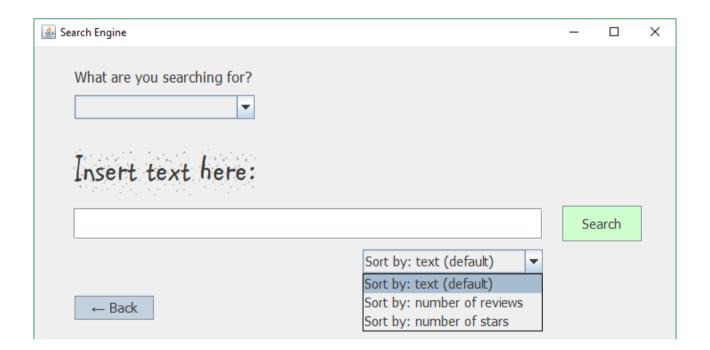
Το πεδίο "What are you searching for?", αναφέρεται στα διαθέσιμα Fields αναζήτησης. Ο χρήστης πρέπει υποχρεωτικά να επιλέξει κάποιο πριν εισάγει το ερώτημα αναζήτησης. Αν ο χρήστης ψάχνει για εστιατόρια, οι επιλογές είναι οι εξής:



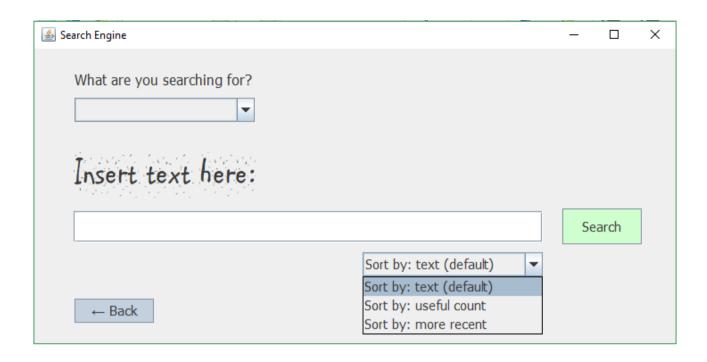
Αν ο χρήστης ψάχνει για κριτικές, οι επιλογές είναι οι εξής:



Το πεδίο "Sort by", αφορά τους τρόπους διάταξης των αποτελεσμάτων. Ο default τρόπος είναι με βάση το κείμενο. Υπάρχουν όμως και άλλοι τρόποι διάταξης. Αν ο χρήστης ψάχνει για εστιατόρια, οι τρόποι αυτοί είναι οι εξής:



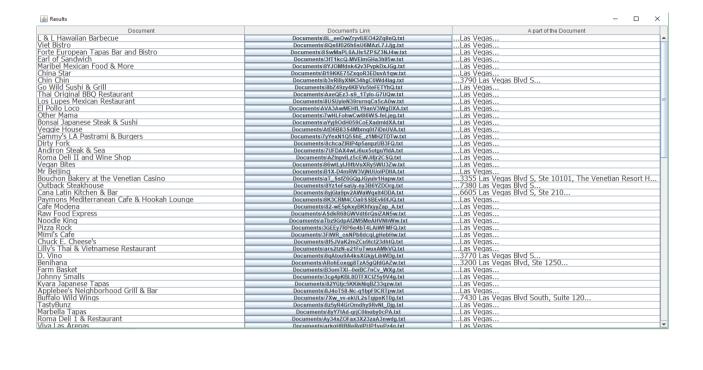
Αν ο χρήστης ψάχνει για κριτικές, οι τρόποι διάταξης είναι οι εξής:



Στη συνέχεια, αφού ο χρήστης εισάγει το ερώτημα αναζήτησης στο πεδίο "Insert text here", θα πρέπει να πατήσει το κουμπί Search, για να πραγματοποιηθεί η αναζήτηση. Αν ο χρήστης θέλει να πραγματοποιήσει αναζήτηση σε 2 Fields θα πρέπει μεταξύ των όρων που θα εισάγει να περιέχονται οι λέξεις OR ή AND.

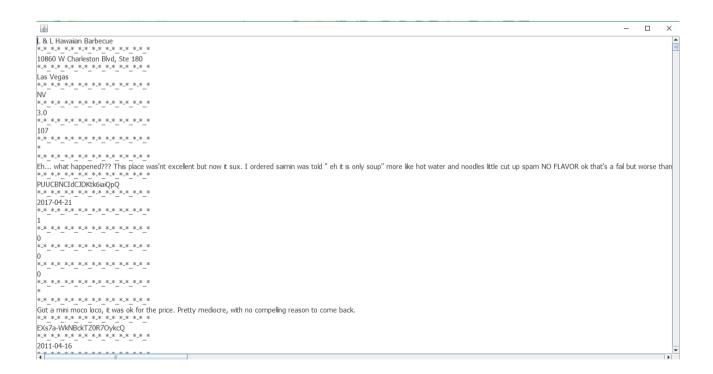
Μόλις τα αποτελέσματα της αναζήτησης είναι έτοιμα, εμφανίζεται ένα νέο παράθυρο, με 3 στήλες και 100 γραμμές. Η πρώτη στήλη αφορά το όνομα του εστιατορίου, η δεύτερη έχει το path του αρχείου που αφορά το συγκεκριμένο εστιατόριο και η τρίτη έχει λίγο από το περιεχόμενο του αρχείου στο οποίο περιέχεται το ερώτημα αναζήτησης. Η κάθε γραμμή είναι και ένα αποτέλεσμα αναζήτησης.

Για παράδειγμα, αν ψάχνουμε για εστιατόρια, στο πεδίο city με ερώτημα αναζήτησης Las Vegas, θα εμφανιστεί:



Με τη βοήθεια της πλαϊνής μπάρας, ο χρήστης μπορεί να δει όλα τα διαθέσιμα αποτελέσματα.

Ένα επιπλέον χαρακτηριστικό αυτού του παραθύρου είναι ότι η δεύτερη στήλη διαθέτει το path του κάθε αρχείου σε μορφή κουμπιού. Οπότε αν ο χρήστης πατήσει το κουμπί θα εμφανιστεί σε ένα νέο παράθυρο το περιεχόμενο του αρχείου, όπως φαίνεται στην εικόνα:



Παραδείγματα

Παρακάτω παρουσιάζονται μερικά παραδείγματα αναζήτησης.

◆ Search Restaurants > review > Sort by: text (default) > essen > Search

(Άνοιγμα τυχαίου εγγράφου)

Search Engine - □ ×

What are you searching for?

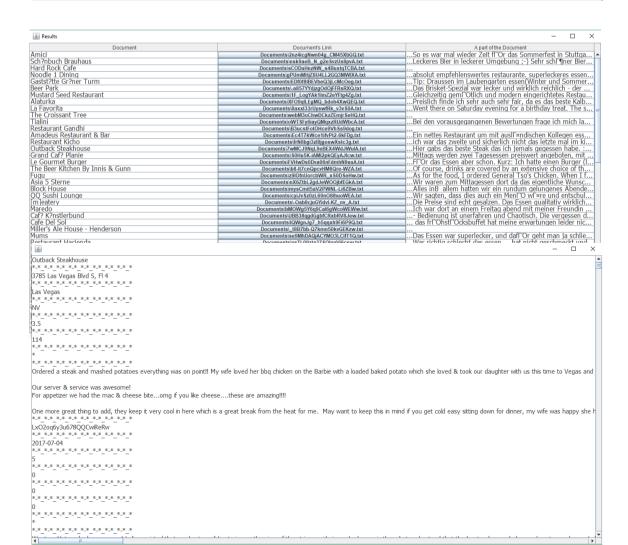
review

Insert text here:

essen

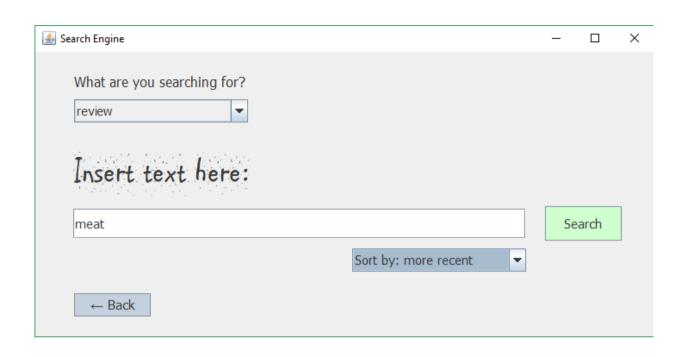
Sort by: text (default)

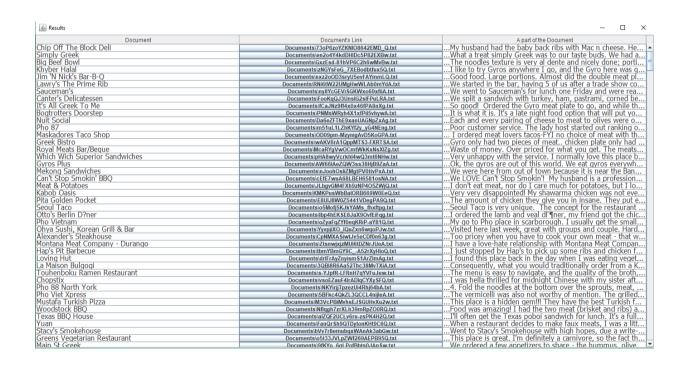
→ Back



♦ Search Reviews > review > Sort by: more recent > meat > Search

(Άνοιγμα τυχαίου εγγράφου)

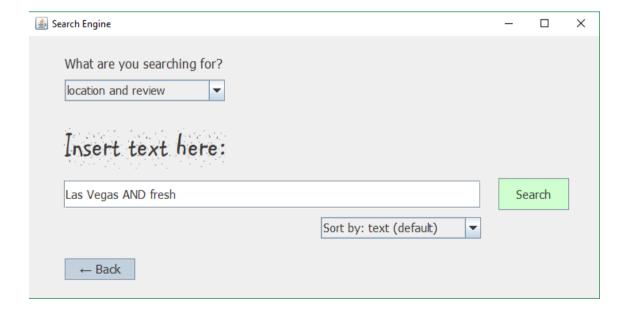




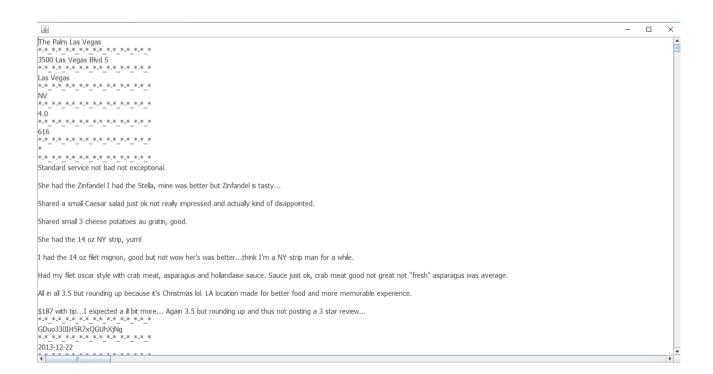


◆ Search Restaurants > location and review > Sort by: text (default)> Las Vegas AND fresh > Search

(Άνοιγμα τυχαίου εγγράφου)

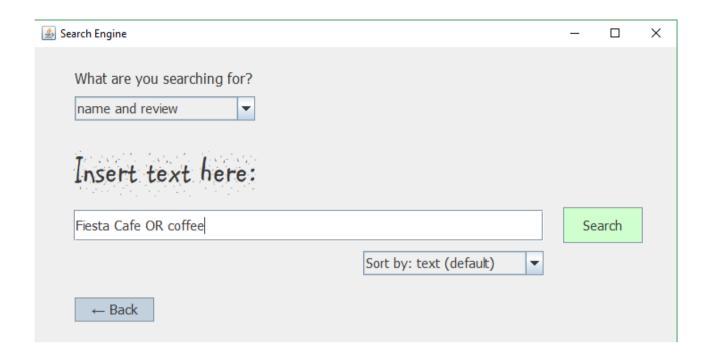


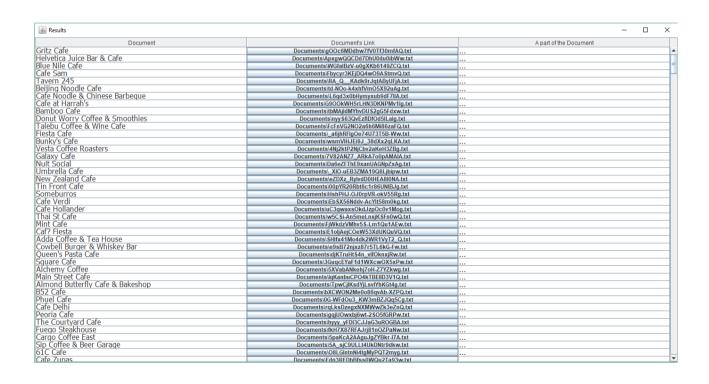
Document	Document's Link	A part of the Document
not Brasserie	Documents\5EluCcoHY2evAhFMBV6G6g.txt	
ake Shack	Documents\QJatAcxYgK1Zp9BRZMAx7g.txt	
Segundo Sol	Documents\Ri-7ymdw8aNZBRqGRAjR3Q.txt	
GM Grand Buffet	Documents\-U7tvCtaraTQ9b0zBhpBMA.txt	
?a Maria Tamales Restaurant	Documents\\yrDnrvUKpndj5zlv1KvGQ.txt	
Islitas Mariscos	Documents\E6DmUAHfSRCEakvknot-zA.txt	
ret Pizza	Documents\iCQpiavjjPzJ5_3gPD5Ebg.txt	
's New York Pizza	Documents\64o nYvzlKwHFOSYSEm-qw.txt	
e at Harrah's	Documents\G9OOkWH5rLHN3DKNPMv1lg.txt	
e Palm Las Vegas	Documents\w12x7UB\$\$Rsdbp3ZqguG3Q.txt	
B: Kerry's Gourmet Burgers	Documents\loSgYD9dHOsjjm7QCly5qw.txt	
Provencal	Documents\78TC3sZSYBzBsSJ0z5pyhw.txt	
rd Rock Cafe	Documents\1CR2ddUcjYYwRd5JLtiRrw.txt	
ising Cane's	Documents\3vxlqhm2jCVjkUWtmDvTsw.txt	
e Garden Grill	Documents\9Jc3W0aR9Xf2gcHl0rEXsw.txt	
er Park	Documents\-a857YYdjzgOdOjFFRsRXQ.txt	
X	Documents\IQ9VmIWvM6v-rONPkYX8aQ.txt	
mphis Championship Barbecue	Documents\LnRn6ECHdqTeoFF0c8FKYw.txt	
Noodle by Jos? Andr?s	Documents\CVRTly4ag_q5ua0R3iqrVw.txt	
na Lounge	Documents\UN_6ebslv5cxdDmZAX3ivQ.txt	
rendipity 3	Documents\eoHdUeQDNgQ6WYEnP2aiRw.txt	
s Artistes Steakhouse	Documents\l_alzz-3iKA3o7G6S4A30w.txt	
tsuya	Documents\RSOinkOUpxm0mGw1IWr4Xw.txt	
s Boulangerie	Documents\RycZOiohghol0Ssg2Qqgqw.txt	
berto's Taco Shop	Documents\-X_w25LQmz6SDcJlFyAPJw.txt	
gar & Ice	Documents\L-zHt7gZITG1z8I7TZ4dUg.txt	
niel Boulud Brasserie	Documents\qmh6zxtJ8C8-YiUPv7ySlw.txt	
icos El Gordo	Documents\RwMLuOkImBlqqYj4SSKSPg.txt	
T - Tacos & Tequila	Documents\MAuSh7NKoTCQ4BumoayA.txt	
e Capital Grille	Documents\Rd11Bosr8JkfoUhZLzpuxg.txt	
rbone	Documents\vZLAmpN-hTl3GCT9hIHVUA.txt	
ada	Documents\G-5kEa6E6PD5fkBRuA7k9Q.txt	
vier's	Documents\Xj7DlGRHEchJ-VVdlSazQQ.txt	
o Tuscan Grille	Documents\UYh1N1xQhOh-a7nX92xPzQ.txt	
ng's Sausage	Documents\7RR1rbyEKKo2G-RACqAfoA.txt	
nsi	Documents\bK_eTLw_uEuEyMzdZlKN6A.txt	
Smacking Foodie Tours	Documents\iBPyahdJRP5y0t25fF2W9w.txt	
Boiler Room	Documents_hML_SleB40KkUrMHOZQ.txt	
ckie's Diner	Documents\l8j1gHLKNJVN4dNVUOsedw.txt	
rgil's Real Barbecue - Las Vegas	Documents\zaxK3ea92mvEBRy-jttngg.txt	
ncon Criollo	Documents\EGI8uU1uf0msVtu8XDrNIw.txt	
ombay Indian Cuisine	Documents\GMBvBDKgOT51xDTpGVphBw.txt	



◆ Search Reviews > review > Sort by: more recent > meat > Search

(Άνοιγμα τυχαίου εγγράφου)





	- □ ×
Talebu Coffee & Wine Cafe	
** * * * * * * * * * * * * * * *	
2095 N Alma School Rd, Ste 2	
* * * * * * * * * * * * * * * * * * * *	
Chandler	
** ** ** ** ** ** ** ** **	
AZ	
4.0 ** ** ** ** ** ** ** ** ** **	
120	
** ** ** ** ** ** ** **	
*	
** * * * * * * * * * * * * * * * *	
This place is very cosy, comfortable, the staff are extremely nice and the coffee is great. Nice place to study or work with slow ambient music playing in the background.	
* * * * * * * * * * * * * * * * * * * *	
Y6SpbQd-ZPWdoQ1J0Q4NGg	
** * * * * * * * * * * * * * * *	
2016-01-25 ** ** * * * * * * * * * * * * * * * *	
D #.* *.* *.* *.* *.* *.* *.* *.* *.*	
U ** *.* *.* *.* *.* *.* *.* *.* *.* *.*	
** ** ** ** ** ** **	
** ** ** ** ** ** ** ** *	
*	
** ** ** ** ** ** ** ** ** *	
3-18-13 Visited this place for the first time today with my daughter. I found them on groupon. I ordered an iced Mexican mocha, never heard of that drink and decided to try it out. It	was pretty good, differ
ewt.3pKy.vXultmFZoUczw ** ** ** ** ** ** ** ** ** **	
2013-03-18	
2/11-70-10	
)