Université d’Avignon TP E-COMMERCE   M TORRES   2020.octobre

Extraction de mots clés – Travaillez en binômes ou individuellement

Dans ce TP nous allons faire une modélisation simple d’extraction de mots-clé à partir de textes scientifiques. Nous avons collecté un certain nombre d’articles scientifiques, venant de plusieurs domaines : informatique, médecine, agronomie, physique, chimie, etc.

Ce sont des papiers dont on a extrait leur texte, puis nous les avons segmenté afin d’avoir une phrase par ligne, avec des metadonnées (title, auteur, date, acronymes, etc). Le champ KEYWORDS est en rouge. En jaune ce sont des formules mathématiques qu’il faudra virer :

```
TYPE   Letter
DOI    http://doi.org/10.1088/2399-6528/aa8215
JOURNAL       Journal of Physics Communications
DATE   01/09/2017
AUTHOR  Wang, Luyang;
ADDRESS       Institute for Advanced Study, Tsinghua University , Beijing, 100084, People's Republic of China
TITLE  Superconductivity in a two-dimensional repulsive Rashba_gas at low electron density
ACRONYMS      Total Density_of States (DOS) ; Renormalization Group (RG) ; Time Reversal Symmetry (TRS) ;
KEYWORDS       superconductivity; low density Rashba_gas; renormalization group; collective modes; Majorana zero modes; repulsive interaction
ABSTRACT      We study the superconducting instability and the resulting superconducting states in a two-dimensional repulsive Fermi gas with Rashba spin-orbit coupling at low electron density (namely the Fermi energy EF is lower than the energy ER of the Dirac point induced by Rashba coupling).
ABSTRACT      We find that superconductivity is enhanced as the dimensionless Fermi energy ϵ F ( ϵ F ≡ E F / E R ) decreases, due to two reasons.
ABSTRACT      First, the density of states at ϵ F increases as 1 / ϵ F .
ABSTRACT      Second, the particle-hole bubble becomes more anisotropic, resulting in an increasing effective attraction.
ABSTRACT      More importantly, once a sufficiently large Zeeman coupling is applied to the superconducting state, the Chern_number can be tuned to be ±1 and Majorana zero modes exist in the vortex cores.
__SECTION__ Intorduction
Despite an effect originating from relativity, spin-orbit coupling (SOC) has found its way into nonrelativistic physics.
In condensed matter physics, novel systems with SOC playing a significant role are found recently, such as topological insulators [1, 2], two-dimensional (2D) Rashba_gases at interfaces of oxides [3, 4], Weyl semimetals [5] and SOC-induced Mott insulators [6] and other states in 5d series [7]; while in ultracold quantum gases, although atoms are neutral, synthetic SOC can be generated by atom-light interaction (see [8, 9] for review).
Turning to superconductivity, non-centrosymmetric superconductors, where SOC mixes spin singlet and triplet pairings, have been extensively studied [10-14]; and in 2D, superconductivity related to SOC was observed at oxide interfaces [4, 15].
Here, we study a 2D repulsive gas with Rashba SOC at low density.
The single-particle Hamiltonian is 1 H = k 2 2 m + α R ( σ and #x020D7; × k ) · n ^ , where m is the effective mass, α R characterizes the strength of Rashba SOC, σ and #x020D7; 's components are Pauli matrices, and n ^ is the direction normal to the 2D system.
By a unitary transformation to helicity basis, one finds the dispersion 2 E k Λ = ( k - Λ k R ) 2 2 m , where Λ = ± 1 is the helicity and k R = m α R is the Rashba momentum.
(We have shifted the energy by k R 2 / ( 2 m ) , which will be compensated by the shift of the Fermi energy.)
The spin degeneracy is lifted, resulting in two bands touching at a Dirac point.
In this system, the competition between the three energy scales-the Fermi energy EF, Coulomb repulsion and the 'Rashba energy' E R = k R 2 / ( 2 m ) -determines the system's phases.
We define the dimensionless Fermi energy by ϵ F = E F / E R .
```

The correction from the particle–hole bubble (shown in figure 6 (a)) is \(< tex-math > < ?CDATA \tfrac{{u}^{2}}{{2}^{6}}({\rm{\Pi }}({\bf{k}},{\bf{k}}^{\prime})-{\rm{\Pi }}(-{\bf{k}},{\bf{k}}^{\prime}))? > < /tex-math > \)u226(π(k,k')−π(−k,k')) , where the dielectric function is < tex-math > < ?CDATA \begin{eqnarray}{\rm{\Pi }}({\bf{k}},{\bf{k}}^{\prime})=\displaystyle \sum _{\alpha \beta }\displaystyle \int \displaystyle \frac{{{\rm{d}}}^{2}{\bf{p}}}{{(2\pi)}^{2}}\displaystyle \frac{{n}_{{\rm{F}}}({E}_{{\bf{p}}\alpha })-{n}_{{\rm{F}}}({E}_{{\bf{p}}+{\bf{k}}-{\bf{k}}^{\prime} \beta })}{{E}_{{\bf{p}}\alpha }-{E}_{{\bf{p}}+{\bf{k}}-{\bf{k}}^{\prime} \beta }}\times {F}_{\alpha \beta }({\bf{k}},{\bf{k}}^{\prime} ,{\bf{p}}),\end{eqnarray}? > < /tex-math > from : <mml:math> <mml:mrow> <mml:mrow> <mml:mi mathvariant="normal">π</mml:mi> </mml:mrow> <mml:mo stretchy="false">(</mml:mo> <mml:mrow> <mml:mi mathvariant="bold">k</mml:mi> </mml:mrow> <mml:mo>,</mml:mo> <mml:mrow> <mml:mi mathvariant="bold">k</mml:mi> </mml:mrow> <mml:mo accent="false">'</mml:mo> <mml:mo stretchy="false">)</mml:mo> <mml:mo>=</mml:mo> <mml:mrow> ... </mml:mrow> <mml:mo accent="false">'</mml:mo><mml:mo>,</mml:mo> <mml:mrow> <mml:mi mathvariant="bold">p</mml:mi> </mml:mrow> <mml:mo stretchy="false">)</mml:mo> <mml:mo>,</mml:mo> </mml:mrow> </math> where < tex-math > < ?CDATA \begin{eqnarray}{F}_{\alpha \beta }({\bf{k}},{\bf{k}}^{\prime} ,{\bf{p}})=(\alpha {{\rm{e}}}^{-{\rm{i}}{\theta}_{{\bf{p}}}}-{{\rm{e}}}^{-{\rm{i}}{\theta }_{{\bf{k}}}})({{\rm{e}}}^{{\rm{i}}{\theta }_{{\bf{k}}^{\prime} }}-\beta {{\rm{e}}}^{{\rm{i}}{\theta }_{{\bf{p}}-{\bf{k}}^{\prime} +{\bf{k}}}})\times (\beta {{\rm{e}}}^{-{\rm{i}}{\theta }_{{\bf{p}}-{\bf{k}}^{\prime} +{\bf{k}}}}-{{\rm{e}}}^{-{\rm{i}}{\theta }_{-{\bf{k}}}})({{\rm{e}}}^{{\rm{i}}{\theta }_{-{\bf{k}}^{\prime} }}-\alpha {{\rm{e}}}^{{\rm{i}}{\theta }_{{\bf{p}}}}).\end{eqnarray}?> </tex-math> and the condition : <mml:math> <mml:mrow> <mml:msub> <mml:mrow> <mml:mi>F</mml:mi> </mml:mrow> <mml:mrow> <mml:mi>α</mml:mi> <mml:mi>β</mml:mi> </mml:mrow> </mml:msub> … </mml:mrow> <mml:msub> <mml:mrow> <mml:mi>θ</mml:mi> </mml:mrow> <mml:mrow> <mml:mi mathvariant="bold">p</mml:mi> </mml:mrow> </mml:msub> </mml:mrow> </mml:msup> <mml:mo stretchy="false">)</mml:mo> <mml:mi>.</mml:mi> </mml:mrow> </math> Since Cooper pairs are expected to form between electrons near the Fermi surfaces, \(< tex-math > < ?CDATA {\bf{k}}? > < /tex-math > \)k and \(< tex-math > < ?CDATA {\bf{k}}^{\prime} ? > < /tex-math > \)k' are restricted to be at Fermi surface μ and Λ , respectively.

Straightforward calculations show that \(< tex-math > < ?CDATA {\rm{\Pi }}({\bf{k}},{\bf{k}}^{\prime})? > < /tex-math > \)π(k,k') can be written in the form < tex-math > < ?CDATA \begin{eqnarray}{\rm{\Pi }}({\bf{k}},{\bf{k}}^{\prime})={{\rm{e}}}^{{\rm{i}}\phi }2m{{\rm{\Lambda }}}_{\mu \lambda }({\epsilon }_{{\rm{F}}},\cos \phi),\end{eqnarray}? > < /tex-math > Λ(ϵF,cosφ) is a real function that depends on the dimensionless Fermi energy \(< tex-math > < ?CDATA {\epsilon }_{{\rm{F}}}? > < /tex-math > \)ϵF , but not on E and E independently.

Then the renormalized coupling appearing in equation (5) reads < tex-math > < ?CDATA \begin{eqnarray}{V}_{\mu \lambda }^{{\rm{r}}}({j}_{z}))=\displaystyle \frac{{u}^{2}m}{{2}^{5}}{V}_{\mu \lambda }^{({j}_{z})}+...,\end{eqnarray}? > < /tex-math > .

The functions \(< tex-math > < ?CDATA {{\rm{\Lambda }}}_{\mu \lambda }^{(S)}({\epsilon }_{{\rm{F}}},\cos \phi)? > < /tex-math > \)ΛμΛ(S)(ϵF,cosφ) are plotted in figure 7 . At \(< tex-math > < ?CDATA {\epsilon }_{{\rm{F}}}\to {1}^{-}? > < /tex-math > \)ϵF→1− , \(< tex-math > < ?CDATA {{\rm{\Lambda }}}_{++}^{(S)}? > < /tex-math > \)Λ++(S) and changes sign due to the change of the helicity of the inner Fermi surface.

Clearly, the functions depend more strongly on φ at smaller \(< tex-math > < ?CDATA {\epsilon }_{{\rm{F}}}? > < /tex-math > \)ϵF .

Up to the fourth order of u , there is only one term that satisfies two conditions: (i) being finite in nonzero angular momentum channel; and (ii) having a logarithmic divergence \(< tex-math > < ?CDATA \mathrm{ln}(A/{\rm{\Omega }})? > < /tex-math > \)ln(A/Ω) , which may give rise to an instability.

This term is shown in figure 6 (b).

Finally, at ϵ F ≳ 1 , jz becomes 2.

The superconductivity predominantly resides on the outer Fermi surface, and interband coupling induces a small gap on the inner Fermi surface.

The superconducting state breaks time reversal symmetry (TRS), and both Fermi surfaces are fully gapped.

Figure 1. (a) The dispersion relation (fixing ky = 0), with the helicity labeled.

...

CORPUS
Il y a un répertoire 10_JPCO qui contient 10 fichiers txt segméntés : il va a servir à mettre à point vos algorithmes
Le test sera effectué sur le répertoire 12_TEST_JPCO
Vous trouverez le fichier de stopwords en anglais : **fonctionnels_en.txt**

Jetez un coup d'œil aux fichiers contenus dans le répertoire 10_JPCO avec un éditeur de texte (`geny, bluefish, gedit`).

ALGORITHME DE BASE
Une première approximation pour extraire les mots clés.:

1/ Garder le Titre, le Résumé et le corps (sections). Ne pas garder le metadonnée KEYWORDS : il servira a EVALUER vos algorithmes.

2/ Virer les équations : balises `<tex-math> … </tex-math>` et `<mml:math>` … `</math>`

3/ Calculer des listes de n-grammes (1 grammes, 2-grammes, 3-grammes…) et les trier par leur fréquence. Filtrer et garder 5 ou 6 n-grammes les plus significatifs comme keywords.

4/ Calculer la performance. On peut calculer la précision de votre algorithme en utilisant :

Précision = Réponses correctes / Total de réponses

Où Réponses correctes = keywords correctes trouvés (vos keywords générées égales aux keywords de l'auteur dans le metadonnée KEYWORDS) ; Total de réponses = Nb de keywords de l'auteur.
Attention a évaluer les keywords en minuscules et peut être sans espaces, afin de gommer leurs différences de caractères.

CONSIGNES

Bien sur que vous pouvez consulter internet pour vous aider. Vous trouverez RAKE par exemple :
https://github.com/aneesha/RAKE/blob/master/rake.py
https://pypi.org/project/rake-nltk/
https://medium.com/datadriveninvestor/rake-rapid-automatic-keyword-extraction-algorithm-f4ec17b2886c
Mais l'objectif est de comprendre et de programmer vos propres codes TAL

Langage de programmation : celui qui vous voulez mais 2 pts de plus si perl est utilisé. Codé autodocumenté (commentaires pertinents)
Rapport : bref description de votre algorithme dans 1 ou 2 pages, avec l'évaluation de performance. +1 pt si code Latex
Rendu : le 8/janvier 2021 avant minuit

Bon travail !